



CO-TRANSCRIPTIONAL SPLICING IN TWO YEASTS

DISSERTATION

ZUR ERLANGUNG DES AKADEMISCHEN GRADES

Doctor rerum naturalium
(Dr. rer. nat.)

VORGELEGT

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden

VON

Herzel, Lydia

geboren am 22. Januar 1986 in Berlin (Deutschland)

GUTACHTER:

Prof. Dr. Stephan Grill, BIOTEC, Technische Universität Dresden

Prof. Karla M. Neugebauer, Ph.D., Yale University

BETREUER:

Prof. Karla M. Neugebauer, Ph.D., Yale University

EINGEREICHT AM:

04. Juni 2015

VERTEIDIGT AM:

10. September 2015

CO-TRANSCRIPTIONAL SPLICING IN TWO YEASTS

LYDIA HERZEL

Lydia Herzel: *Co-transcriptional splicing in two yeasts*
(c) June 2015

E-mail: lydia.herzel@outlook.com

SUMMARY

Cellular function and physiology are largely established through regulated gene expression. The first step in gene expression, transcription of the genomic DNA into RNA, is a process that is highly aligned at the levels of initiation, elongation and termination. In eukaryotes, protein-coding genes are exclusively transcribed by RNA polymerase II (Pol II). Upon transcription of the first 15-20 nucleotides (nt), the emerging nascent RNA 5' end is modified with a 7-methylguanosyl cap. This is one of several RNA modifications and processing steps that take place during transcription, i. e. co-transcriptionally. For example, protein-coding sequences (exons) are often disrupted by non-coding sequences (introns) that are removed by RNA splicing. The two transesterification reactions required for RNA splicing are catalyzed through the action of a large macromolecular machine, the spliceosome. Several non-coding small nuclear RNAs (snRNAs) and proteins form functional spliceosomal subcomplexes, termed snRNPs. Sequentially with intron synthesis different snRNPs recognize sequence elements within introns, first the 5' splice site (5' SS) at the intron start, then the branch-point and at the end the 3' splice site (3' SS). Multiple conformational changes and concerted assembly steps lead to formation of the active spliceosome, cleavage of the exon-intron junction, intron lariat formation and finally exon-exon ligation with cleavage of the 3' intron-exon junction. Estimates on pre-mRNA splicing duration range from 15 sec to several minutes or, in terms of distance relative to the 3' SS, the earliest detected splicing events were 500 nt downstream of the 3' SS. However, the use of indirect assays, model genes and transcription induction/blocking leave the question of when pre-mRNA splicing of endogenous transcripts occurs unanswered.

In recent years, global studies concluded that the majority of introns are removed during the course of transcription. In principal, co-transcriptional splicing reduces the need for post-transcriptional processing of the pre-mRNA. This could allow for quicker transcriptional responses to stimuli and optimal coordination between the different steps. In order to gain insight into how pre-mRNA splicing might be functionally linked to transcription, I wanted to determine when co-transcriptional splicing occurs, how transcripts with multiple introns are spliced and if and how the transcription termination process is influenced by pre-mRNA splicing.

I chose two yeast species, *S. cerevisiae* and *S. pombe*, to study co-transcriptional splicing. Small genomes, short genes and introns, but very different number of intron-containing genes and multi-intron genes in *S. pombe*, made the combination of both model organisms a promising system to study by next-generation sequencing and to learn about co-transcriptional splicing in a broad context with applicability to other species. I used nascent RNA-Seq to characterize co-transcriptional splicing in *S. pombe* and developed two strategies to obtain single-molecule information on co-transcriptional splicing of endogenous genes:

(1) with paired-end short read sequencing, I obtained the 3' nascent transcript ends, which reflect the position of Pol II molecules during transcription, and the splicing status of the nascent RNAs. This is detected by sequencing the exon-intron or exon-exon junctions of the transcripts. Thus, this strategy links Pol II position with intron splicing of nascent RNA. The increase in the fraction of spliced transcripts with further distance from the intron end provides valuable information on when co-transcriptional splicing occurs.

(2) with Pacific Biosciences sequencing (PacBio) of full-length nascent RNA, it is possible to determine the splicing pattern of transcripts with multiple introns, e. g. sequentially with transcription or also non-sequentially. Part of transcription termination is cleavage of the nascent transcript at the polyA site. The splicing status of cleaved and non-cleaved transcripts can provide insights into links between splicing and transcription termination and can be obtained from PacBio data.

I found that co-transcriptional splicing in *S. pombe* is similarly prevalent to other species and that most introns are removed co-transcriptionally. Co-transcriptional splicing levels are dependent on intron position, adjacent exon length, and GC-content, but not splice site sequence. A high level of co-transcriptional splicing is correlated with high gene expression. In addition, I identified low abundance circular RNAs in intron-containing, as well as intronless genes, which could be side-products of RNA transcription and splicing.

The analysis of co-transcriptional splicing patterns of 88 endogenous *S. cerevisiae* genes showed that the majority of intron splicing occurs within 100 nt downstream of the 3' SS. Saturation levels vary, and confirm results of a previous study. The onset of splicing is very close to the transcribing polymerase (within 27 nt) and implies that spliceosome assembly and conformational rearrangements must be completed immediately upon synthesis of the 3' SS.

For *S. pombe* genes with multiple introns, most detected transcripts were completely spliced or completely unspliced. A smaller fraction showed partial splicing with the first intron being most often not spliced. Close to the polyA site, most transcripts were spliced, however uncleaved transcripts were often completely unspliced. This suggests a beneficial influence of pre-mRNA splicing for efficient transcript termination.

Overall, sequencing of nascent RNA with the two strategies developed in this work offers significant potential for the analysis of co-transcriptional splicing, transcription termination and also RNA polymerase pausing by profiling nascent 3' ends. I could define the position of pre-mRNA splicing during the process of transcription and provide evidence for fast and efficient co-transcriptional splicing in *S. cerevisiae* and *S. pombe*, which is associated with highly expressed genes in both organisms. Differences in *S. pombe* co-transcriptional splicing could be linked to gene architecture features, like intron position, GC-content and exon length.

PUBLICATIONS

During the course of this thesis my research has contributed to the following publications:

Lydia Herzel and Karla M. Neugebauer. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods*, April 2015.

PMID: 25929182 ([PubMed](#))

Mattia Brugiolo*, Lydia Herzel*, and Karla M. Neugebauer. Counting on co-transcriptional splicing. *F1000 Prime Reports*, 5(9):9, April 2013. *Authors contributed equally and are listed alphabetically.

PMID: 23638305 ([PubMed](#))

ACKNOWLEDGEMENTS

Many people have supported me in my work and accompanied on my way. My sincere thanks go to ...

- ... *Karla Neugebauer* for her marvelous supervision, support and the possibility to explore the scientific world with her at the MPI-CBG and Yale.
- ... *Fernando Carrillo Oesterreich* for the many scientific discussions, advise and often short, but very helpful comments in many phases of this thesis work.
- ... *Korinna Straube* for help with carrying out large scale SMIT experiments, creating a joyful lab atmosphere and also for allowing me to win Skat matches occasionally.
- ... my thesis advisory committee - *Stephan Grill, Gerhard Roedel, Michael Hiller* - for discussions and input in committee meetings.
- ... *Miguel Coelho, Julien Berro, Megan King, Ronan Fernandez* for helpful discussions on *S. pombe* specific biology and methods and for providing strains for *S. pombe* work.
- ... *David Query and his lab* for a fun day discussing SMIT in NYC and for providing *S. cerevisiae* and *S. pombe* strains for future experiments.
- ... *Kaya Bilguvar, Andreas Dahl, Jamie Miller and Guilin Wang* for taking care of the various deep sequencing and PacBio samples I submitted to the YCGA and Sequencing facility, Biotec Dresden.
- ... *Holger Brandl* for introducing me to and providing tips for the use of the Bioinfoserver and the Madmax cluster at the MPI-CBG.
- ... *Tara Alpert and Jeremy Schofield* for making me enjoy PhD student rotation supervisions.
- ... *Mattia Brugiolo, Danielle Krasner, Martin Machyna, Jonathan Rodenfels, David Stanek, Vladimir Despic* for helpful discussions in lab and lab meetings and fun lunch breaks.
- ... *Patricia Heyn, Michaela Mueller-McNicoll* for their scientific advise and mentorship in the first half of my PhD.
- ... *my parents, my brother and friends from Tuebingen, Dresden, Berlin and New Haven* for their strong support, frequent visits, distractions through sports, parties and traveling, and special thanks go to *my father Hans-Peter Herzel* for observing, commenting and discussing scientific progress with me.
- ... *Michael Weber* for his L^AT_EX-support, incredible company and impressive eye for details and the global picture in Science and all other aspects of life.

CONTENTS

i	INTRODUCTION	1
1	INTRODUCTION	3
1.1	Transcription and pre-mRNA splicing of protein-coding genes . . .	3
1.2	Next-generation sequencing	7
1.3	RNA sequencing to study co-transcriptional splicing	9
1.4	Two yeasts and their gene architecture	12
2	THESIS AIM AND OVERVIEW	15
ii	RESULTS	17
3	QUANTIFICATION OF CO-TRANSCRIPTIONAL SPLICING	19
3.1	Strategies to quantify co-transcriptional splicing	19
3.1.1	Nascent RNA-Seq in previous studies	19
3.1.2	Quantification of co-transcriptional pre-mRNA splicing in previous studies	19
3.1.3	Transcription-associated aspects of quantification	22
3.1.4	Comparison between splicing measures	23
3.1.5	Applicability to yeast pre-mRNA splicing	24
3.2	Quantification of co-transcriptional splicing in <i>S. pombe</i>	24
3.2.1	Preparation of nascent RNA from <i>S. pombe</i> chromatin	24
3.2.2	<i>S. pombe</i> nascent and mRNA-Seq	26
3.2.3	Global pre-mRNA splicing estimates from chromatin-associated RNA and mRNA	29
3.2.4	Patterns of co- and post-transcriptional splicing	31
4	CO-TRANSCRIPTIONAL SPLICING IN <i>s. pombe</i>	33
4.1	Gene function and expression	33
4.2	Gene architecture features of differentially spliced introns	36
4.3	Coupling of intron splicing to other processing events	41
4.3.1	Appearance of circular RNAs	41
4.3.2	Long read sequencing to link co-transcriptional RNA pro- cessing events	44
4.3.3	Order of intron removal	47
4.3.4	PolyA site cleavage and co-transcriptional splicing	49
4.4	Multiple links between pre-mRNA splicing, gene architecture and expression	51
5	KINETICS OF CO-TRANSCRIPTIONAL SPLICING	53
5.1	A method for Single Molecule Intron Tracking along with tran- scription	53
5.1.1	3' DNA adaptor ligation to preserve Pol II position during transcription	54
5.1.2	SMIT PCR and library to identify splicing position	55
5.2	The position of co-transcriptional splicing in <i>S. cerevisiae</i> genes	58

5.2.1	Common and unique characteristics of co-transcriptional splicing in individual genes	58
5.2.2	Frequent and rapid co-transcriptional splicing for most assayed genes	59
5.3	Fast co-transcriptional splicing is widespread	59
iii	DISCUSSION	63
6	CO-TRANSCRIPTIONAL SPLICING	65
6.1	Global co-transcriptional splicing levels, kinetics and transcriptional pausing	65
6.2	General and <i>S. pombe</i> specific intron splicing characteristics	68
7	STRATEGIES OF SEQUENCING NASCENT RNA	73
7.1	Quantification of co-transcriptional splicing from RNA-Seq data	73
7.1.1	Removal of polyadenylated RNA	74
7.1.2	Removal of ribosomal RNA	75
7.2	Chromatin-associated transcripts from PacBio sequencing	76
7.2.1	snoRNAs in libraries with varying length spectrum	77
7.2.2	pre-mRNA splicing in libraries with varying length spectrum	78
7.3	SMIT as general application to study co-transcriptional RNA processing	79
8	CONCLUDING REMARKS	83
iv	METHODS	85
9	METHODS	87
9.1	Yeast strains	87
9.1.1	<i>S. pombe</i> strains	87
9.1.2	<i>S. cerevisiae</i> strains	87
9.2	Yeast growth	88
9.2.1	<i>S. pombe</i> growth	88
9.2.2	<i>S. cerevisiae</i> growth	88
9.3	Cell harvest	88
9.3.1	Filtration	88
9.3.2	Centrifugation	89
9.4	Cell fractionation and nascent RNA preparation	89
9.4.1	Cell fractionation & RNA extraction	89
9.4.2	Ethanol precipitation	90
9.4.3	DNase treatment	90
9.4.4	Removal of polyA+ RNA	90
9.4.5	Qualitative and quantitative analysis of nucleic acids	90
9.4.6	Removal of rRNA	91
9.4.7	Buffers	91
9.5	Isolation of total RNA	93
9.6	RT-(q)PCR	93
9.6.1	Reverse transcription (RT)	93
9.6.2	qPCR	93
9.6.3	PCR	94
9.7	DNA purification	94

9.8	Protein analysis	94
9.9	RNA-Seq	94
9.10	3' end ligation	95
9.11	PacBio nascent RNA libraries	95
9.12	Single Molecule Intron Tracking (SMIT) libraries	96
9.13	<i>In vitro</i> transcription (IVT)	97
9.14	Mapping of RNA-Seq data	97
9.15	Quantification of intron and exon splicing levels	97
9.16	Analysis of pre-mRNA splicing characteristics	98
9.16.1	GO term analysis	98
9.16.2	Gene expression analysis	98
9.16.3	Gene architecture analysis	99
9.17	PacBio data processing & mapping	99
9.18	SMIT data processing & mapping	100
9.19	SMIT data normalization and splicing analysis	100
V	APPENDIX	103
A	APPENDIX A	105
A.1	Influence of the number of replicates on quantification	105
A.2	Low data cutoff to remove background noise	105
A.3	Correlation of intron-centric approaches between all analyzed samples	110
A.4	Handling alternative splicing	111
B	APPENDIX B	113
B.1	Mapping & correlation of <i>S. pombe</i> RNA-Seq data	113
B.2	Impact of caffeine on <i>S. pombe</i> gene expression and splicing	116
B.3	Expression and intron splicing differences in <i>S. cerevisiae</i> paralogs	118
B.4	Splicing of short internal introns	122
B.5	Sequencing full-length nascent RNA with Pacific Biosciences sequencing	123
B.6	Examples of sequential and non-sequential splicing	127
C	APPENDIX C	129
C.1	Data processing and mapping of SMIT data	129
C.2	Single gene SMIT examples	134
C.3	SMIT is not biased by degraded RNA	137
D	APPENDIX D	139
D.1	Primers <i>S. pombe</i>	139
D.2	Primers for Single Molecule Intron Tracking	141
D.3	Primers for <i>in vitro</i> transcription and transcribed RNA sequences	152
D.4	Primers for PacBio libraries	155
	BIBLIOGRAPHY	157

LIST OF FIGURES

Figure 1	Transcription and co-transcriptional RNA processing . . .	4
Figure 2	Next generation sequencing technology	8
Figure 3	Global co-transcriptional splicing values from nascent RNA sequencing	11
Figure 4	<i>S. cerevisiae</i> and <i>S. pombe</i> distance in evolution and gene architecture	13
Figure 5	Global co-transcriptional splicing values in mouse liver and macrophages	21
Figure 6	Intron pre-mRNA splicing and 5' to 3' abundance gradient	22
Figure 7	Correlation between approaches	23
Figure 8	Enrichment of nascent RNA	25
Figure 9	GO analysis of chromatin and cytoplasmic fraction	28
Figure 10	RNA-Seq of three cellular fractions	29
Figure 11	Splicing of <i>S. pombe</i> introns	30
Figure 12	GO analysis of the four groups with differential splicing patterns	34
Figure 13	Co- and post-transcriptional gene splicing values correlate with mRNA expression	35
Figure 14	Gene expression changes correlate weakly with co-transcriptional splicing changes	36
Figure 15	Differences in intron position and exon length associated with pre-mRNA splicing	39
Figure 16	Differences in SS strength and GC-content associated with pre-mRNA splicing	40
Figure 17	Circular RNA detection and characterization in <i>S. pombe</i> .	43
Figure 18	Long read sequencing library preparation and data processing	46
Figure 19	Connectivity of splicing of multiple introns	48
Figure 20	Co-transcriptional splicing and polyA site cleavage	50
Figure 21	Single Molecule Intron Tracking (SMIT) library overview .	54
Figure 22	Characterization of the 3' end DNA adaptor ligation	55
Figure 23	SMIT PCRs and library	57
Figure 24	Onset and progression of co-transcriptional can be monitored by SMIT	58
Figure 25	Co-transcriptional splicing is efficient and fast	59
Figure 26	Early co-transcriptional splicing for splicing reporter gene	61
Figure 27	Transcription- and splicing factor distribution around the terminal exon pause site	67
Figure 28	Influence of polyA+ RNA and rRNA depletion on splicing quantification	74
Figure 29	Read length considerations in PacBio sequencing	77

Figure 30	Average splicing levels in different PacBio libraries	78
Figure 31	Detection of terminal exon pausing	80
Figure 32	Correlation between replicates	105
Figure 33	Low data cutoff	106
Figure 34	Minimal cutoff for SPI	107
Figure 35	Minimal cutoff for coSI	108
Figure 36	Minimal cutoff for Intron Difference	109
Figure 37	Pearson correlation of SPI and SS ratio and between dif- ferent mouse samples	110
Figure 38	Lower pre-mRNA splicing of alternative introns .	111
Figure 39	Gene expression and pre-mRNA splicing correlation . . .	115
Figure 40	<i>S. pombe</i> cell growth and survival upon Caffeine treatment	116
Figure 41	Splicing and gene expression correlation	117
Figure 42	Intron groups and gene expression changes	118
Figure 43	Intron-containing paralogs in <i>S. cerevisiae</i> differ strongly in intron length, sequence and slightly in splice site con- servation	120
Figure 44	Expression differences between intron-containing paralogs in <i>S. cerevisiae</i> and splice site conservation differences . . .	121
Figure 45	Internal intron splicing and adjacent exon length	122
Figure 46	Pacific Biosciences library design and characterization . .	124
Figure 47	SnoRNA localization and annotation	125
Figure 48	Correlation of PacBio splicing and RNA-Seq SPIs	126
Figure 49	Examples of (non-) sequential splicing	128
Figure 50	SMIT library design, preparation and data processing . . .	132
Figure 51	SMIT replicate correlation and read count per position distribution	133
Figure 52	Examples of endogenous co-transcriptional splicing patterns	135
Figure 53	Correlation of SMIT saturation values, tiling array data and 3' PE-Seq data	136
Figure 54	<i>In vitro</i> analysis of ligation to degraded RNA	137

LIST OF TABLES

Table 1	Genome and gene features	14
Table 2	Mass spectrometry of <i>S. cerevisiae</i> and <i>S. pombe</i> chromatin	27
Table 3	Pre-mRNA splicing and gene architecture	37
Table 4	Gene properties associated with pre-mRNA splicing . . .	69
Table 5	<i>S. pombe</i> strains	87
Table 6	<i>S. cerevisiae</i> strains	87
Table 7	Buffer 1	91
Table 8	Buffer 2, pH 7.6	92
Table 9	Buffer P, pH 5.0	92
Table 10	RNA-Seq mapping	114
Table 11	PacBio sequencing and mapping details	123
Table 12	SMIT raw, processed and mapped read counts (1)	130
Table 13	SMIT raw, processed and mapped read counts (2)	131
Table 14	<i>S. pombe</i> primer sequences	140
Table 15	<i>S. cerevisiae</i> SMIT primer sequences (1)	142
Table 16	<i>S. cerevisiae</i> SMIT primer sequences (2)	143
Table 17	<i>S. cerevisiae</i> SMIT primer sequences (3a)	144
Table 18	<i>S. cerevisiae</i> SMIT primer sequences (3b)	145
Table 19	<i>S. cerevisiae</i> SMIT primer sequences (3c)	146
Table 20	<i>S. cerevisiae</i> SMIT primer sequences (3d)	147
Table 21	<i>S. cerevisiae</i> SMIT primer sequences (3e)	148
Table 22	<i>S. cerevisiae</i> SMIT primer sequences (3f)	149
Table 23	<i>S. cerevisiae</i> SMIT primer sequences (3g)	150
Table 24	<i>S. cerevisiae</i> SMIT primer sequences (4)	151
Table 25	<i>S. cerevisiae</i> SMIT primer sequences (5)	151
Table 26	3' end ligation (1)	153
Table 27	3' end ligation (2)	154
Table 28	PacBio library primers	156

ACRONYMS

3' SS	3' splice site
5' SS	5' splice site
bp	base pair
CCS	circular consensus sequence
coSI	completed splicing index
ds	downstream
EM	electron microscopy
FDR	false discovery rate
FPKM	Fragments Per Kilobase Of Exon Per Million Fragments Mapped
IVT	<i>in vitro</i> transcription
kb	kilobases
Mio	Millions
mRNA	messenger RNA
nRNA	nascent RNA
nt	nucleotide
OD	optical density
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
polyA site	polyadenylation site
Pol II	RNA polymerase II
pre-mRNA	pre-messenger RNA
RNA-Seq	RNA sequencing
rRNA	ribosomal RNA
RT	Reverse transcription
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
SMIT	Single molecule intron tracking

snoRNA	small nucleolar RNA
SPI	splicing per intron
SS	splice site
us	upstream

Part I

INTRODUCTION

INTRODUCTION

1.1 TRANSCRIPTION AND PRE-MRNA SPLICING OF PROTEIN-CODING GENES

Gene expression in eukaryotic cells is a highly regulated, multi-step process that establishes cellular identity and function. Gene expression starts with reading out gene information through transcription by three DNA-dependent RNA polymerases. RNA polymerase I (Pol I) synthesizes the 45S (35S in yeast) ribosomal RNA (rRNA) precursor, which matures into 28S, 18S and 5.8S rRNA. rRNAs make up ~95% of cellular RNAs and serve as structural and functional components of the ribosome [Grummt 1999]. tRNAs, 5S rRNA and other small non-coding RNAs are synthesized by RNA polymerase III (Pol III) [Willis 1993]. The synthesis of protein-coding messenger RNA (mRNA) is exclusively carried out by RNA polymerase II (Pol II) [Kornberg 1999].

The process of transcription can be separated into three main parts: initiation, elongation and termination (Figure 1). By now, it has become clear that all three phases are subject to regulation and consist of multiple tightly connected steps. In protein-coding genes, Pol II is recruited to the promoter region by the basal transcription machinery and co-activators, and a transcription preinitiation complex is formed. With melting of the DNA double-strand the transition to an open initiation complex takes place and initial transcription starts [Cramer 2004]. Initiating Pol II must undergo structural and functional maturation to processively transcribe the genomic information. Thus, several rounds of abortive initiation with synthesis of nascent RNAs shorter than 10 nts often precede final promoter clearance [Dvir 2002]. Entering transcription elongation, a stable ternary complex between the enzyme Pol II, the genomic DNA and the nascent RNA chain is formed. Nucleotides can be paired with the template and are joined processively [Saunders et al. 2006]. Elongation factors modulate progression in transcription elongation [Kwak et al. 2013].

A unique feature of Pol II is the long carboxy-terminal domain (CTD) of its largest subunit, Rpb1. The CTD consists of tandem repeats of seven amino acids, YSPTSPS. The CTD is subject to post-translational modifications throughout the transcription cycle. During initiation it is hypophosphorylated, and over the course of transcription elongation phosphorylated at Serine 2 and Serine 5. Serine 5 phosphorylation peaks early in the transcription cycle, whereas Serine 2 phosphorylation predominates in the body and towards the 3' end of the gene [Mayer et al. 2010]. Other amino acids are also modified throughout transcription elongation, e. g. Serine 7 and Tyrosine 4 [Mayer et al. 2012]. The modification of the CTD affects its conformation and the ability to associate with factors that are involved in transcription elongation, nascent RNA processing and termination [Hsin and Manley 2012].

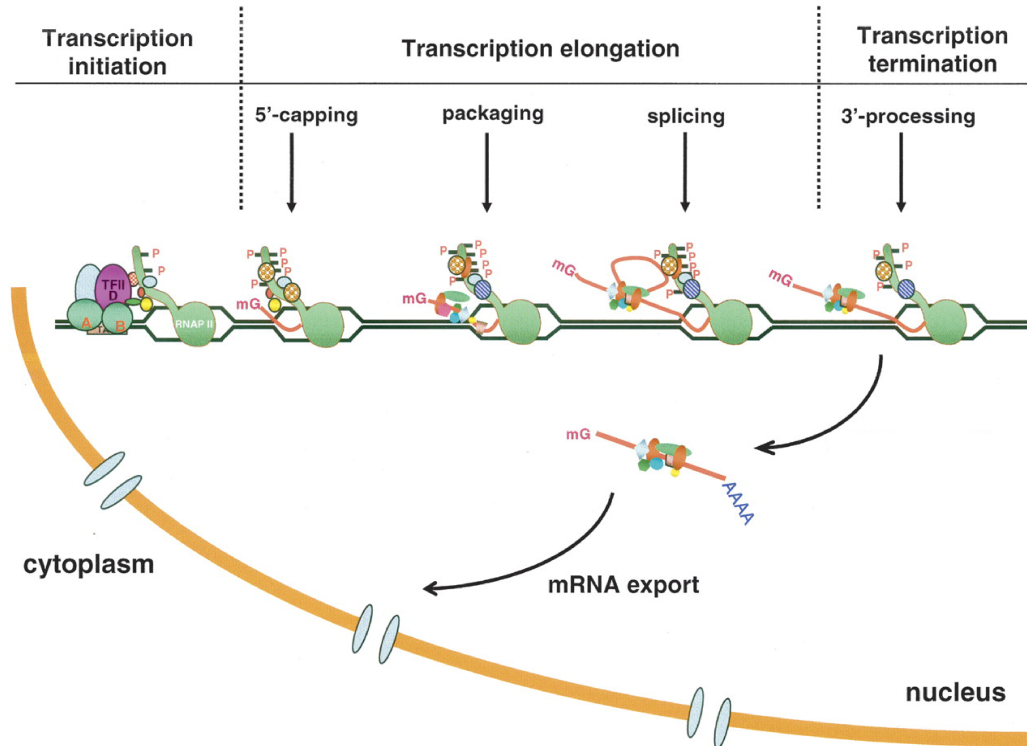


FIGURE 1 Transcription and co-transcriptional RNA processing. Transcription of protein-coding genes by RNA polymerase II (RNAP II, green with tail, the C-terminal domain, CTD) can be divided into 3 phases, transcription initiation, elongation and termination. Transcription initiation consists of (pre)initiation complex formation and promoter clearance. During elongation the stable ternary complex of RNAP II, nascent RNA (orange) and the template DNA (darkgreen) is formed and processive transcription occurs. The 5' end of nascent RNA is capped co-transcriptionally with the 7-methylguanosine cap (mG). The CTD serves as a binding platform for proteins involved in transcription, RNA processing and mRNP (messenger ribonucleoprotein) formation throughout transcription. Various RNA-binding proteins, e.g. splicing factors, export factors or termination factors, associate with the nascent RNA co-transcriptionally (packaging) and also pre-mRNA splicing can take place co-transcriptionally. After co-transcriptional nascent RNA cleavage at the polyA site the transcript 3' end gets polyadenylated (blue As). The mature mRNA and associated proteins forming the mature mRNP can be exported from the nucleus to the cytoplasm, where protein synthesis (translation) can occur. Figure from [Li and Manley 2006].

The third and last phase of transcription is transcription termination, which is triggered through transcription of the A-rich polyA site and recognition of this site by RNA-binding factors of the termination machinery [Proudfoot 2011]. The process of termination includes (a) cleavage of the nascent transcript, (b) addition of ~200 adenine residues, which form the polyA tail of mature mRNA, and (c) degradation of the remaining nascent RNA still associated in the transcription elongation complex and finally (d) termination of transcription [Kuehner et al. 2011].

Genomic DNA is wrapped and packaged around histones, which together form nucleosomes, the repeating units of eukaryotic chromatin. These obstacles must be overcome by the transcription machinery. Nucleosome remodelling factors help disassembling and reforming them as Pol II moves along the gene. Post-translational modifications on amino acids of the histone protein tails, so called chromatin marks, can serve as binding platform for transcription factors and can also change the physical properties of nucleosomes and therefore aid or hinder transcription [Li et al. 2007].

During the course of transcription, the nascent RNA grows in 5' to 3' direction. After ~15 nt have been synthesized, the 5' end of nascent RNA gets covalently modified through the three-step addition of a methylated guanine base [Martinez-Rucobo et al. 2015]. The resulting 7-methylguanosine cap enhances stability of the nascent RNA, promotes transcription, RNA processing and nuclear export of mRNA [Cowling 2010].

Some protein-coding genes in *S. cerevisiae* and most protein-coding genes in higher eukaryotes are disrupted by stretches of non-coding sequences, so-called introns [Berget et al. 1977, Chow et al. 1977]. Those need to be removed before protein synthesis occurs. The intron removal and joining of coding sequences (exons) to form mature mRNA is called pre-mRNA splicing. This process is catalyzed by the protein-RNA machinery, the spliceosome, a macromolecular machine the size and complexity of the ribosome that can already assemble during Pol II transcription [Görnemann et al. 2005, Listerman et al. 2006, Lacadie and Rosbash 2005, Lacadie et al. 2006]. The spliceosome is assembled from small nuclear RNAs (snRNAs) and protein complexes. Subcomplexes are called small nuclear ribonucleoproteins (snRNPs).

Splicing is especially prevalent in higher eukaryotes, where it enhances the informational diversity and functional capacity of a gene. Often, there are multiple splice variants, which can originate from a single gene locus. This variation is called alternative splicing and is distinguished from constitutive splicing, which is not subject to regulation and produces the majority of spliced transcripts. Regulation of alternative splicing is achieved by cis-regulatory elements in introns and exons in combination with various trans-acting factors [Shepard and Hertel 2009]. Mutations leading to mis-regulation of alternative splicing are linked to numerous human diseases, and these mutations can be located in the affected gene or in protein components of the splicing machinery [Singh and Cooper 2012].

Splicing catalysis proceeds via numerous consecutive steps. The 5' splice site (5' SS) is transcribed first and immediately recognized by the U1 snRNP. As

Pol II progresses through the intron, the branchpoint - an adenine residue close to the intron end - and the 3' splice site (3' SS) are recognized by splicing factors. This event triggers the joining of additional spliceosomal components. Multiple conformational rearrangements lead to formation of the active spliceosome. In two catalytic steps the exon-exon ligated product and the excised intron lariat (5' intron end ligated via a 2'-5' phosphodiester bond to branchpoint adenine) are formed [Wahl et al. 2009]. Weak splice sites and a high number of introns per gene facilitate the formation of alternatively spliced transcripts. Splicing catalysis takes place during transcription (co-transcriptionally), like 5' capping, and after transcription termination (post-transcriptionally), when the nascent transcript is already cleaved at the polyA cleavage site and polyadenylated subsequently.

"Co-transcriptionality" of splicing could simply be a coincidence, but it might enable also precise regulation, high efficiency and ensure the correct identity and fate of RNA transcripts. As for 5' capping [Moteki and Price 2002, Martinez-Rucobo et al. 2015], multiple studies provide evidence for a beneficial interaction between the splicing and transcription machinery for gene expression. For example, an intron increases the expression of mouse transgenes [Brinster et al. 1988]. The splice site choice in alternatively spliced genes in metazoans and the splicing efficiency in yeast changes when mutant versions of Pol II are introduced that result in slower or faster transcription elongation [Braberg et al. 2013, Fong et al. 2014, Dujardin et al. 2014]. Furthermore, chromatin marks that correlate with active transcription, e. g. H₃K₄me₃, are also enriched on 5' SSs, while others, e. g. H₃K₃₆me₃, are concentrated over exons [de Almeida et al. 2011]. Splicing inhibition or intron deletion lead to a reduction of such chromatin modifications and transcriptional output [Bieberstein et al. 2012].

The first observation of co-transcriptional splicing was made in highly transcribed *Drosophila* chorion transcripts using electron microscopy [Osheim et al. 1985]. Experiments on specific genes, which were selected because of special properties such as transcriptional inducibility, gene length, and accessibility to light and/or EM microscopy provided further evidence [Pandya-Jones 2011]. In recent years, global studies in multiple organisms and cell types have been performed and it has become clear, that the majority of introns is removed co-transcriptionally [Brugiolo et al. 2013].

The discovery of a dominance of co-transcriptional splicing came as a surprise - especially in the case of *S. cerevisiae*. A previous study concluded that post-transcriptional splicing must be the rule rather than co-transcriptional splicing [Tardiff and Rosbash 2006], because reported average elongation rate and splicing duration (2 kb/min [Mason and Struhl 2005, Marcello 2014], 15-30 sec [Huranová et al. 2010, Martin et al. 2013]) did not allow for co-transcriptional splicing in the time window set by the distance between intron end and gene 3' end (median terminal exon length 582 nt). Till then splicing in yeast was estimated to start 500 nt downstream of the 3' SS [Lacadie et al. 2006]. The identification of splicing associated pausing of Pol II, e. g. terminal exon pausing in highly co-transcriptionally spliced *S. cerevisiae* genes, lead to the conclusion that transcription elongation is not uniform and that modulations of transcription speed might facilitate and allow for co-transcriptional splicing [Carrillo Oesterreich

et al. 2010]. However, it was still unclear when co-transcriptional splicing can take place during transcription in pausing genes, but also in non-pausing genes with lower co-transcriptional splicing.

The time required for pre-mRNA splicing can be measured in two ways: One possibility is as time in second to minute estimates from microscopy data analysis, e. g. FRAP [Huranová *et al.* 2010], live-cell imaging [Coulon *et al.* 2014, Martin *et al.* 2013], or after gene induction, e. g. by RT-qPCR [Singh and Padgett 2009, Alexander *et al.* 2010, Aitken *et al.* 2011]. The second possibility is as distance in nucleotides to kilobases relative to the 3' SS [Görnemann *et al.* 2005, Lacadie and Rosbash 2005, Lacadie *et al.* 2006, Tardiff and Rosbash 2006]. Estimates from previous studies range from 15 sec to several min and from 500 nt to several kb. However, those experiments have been carried out using model genes, indirect readouts like splicing factor ChIP (immunoprecipitation of splicing factors and qPCR) or transcription induction/repression, which can lead to splicing independent changes in RNA stability [Haimovich *et al.* 2013].

The commitment to pre-mRNA splicing is achieved by splicing factor binding to splice sites and other cis-regulatory elements. This and the time it takes to remove an intron have further implications on which introns are removed in transcripts with several introns and how alternative isoforms arise [Kornblihtt 2015]. Data from human and mouse cells suggest that pre-mRNA splicing happens sequentially in the direction of transcription [de la Mata *et al.* 2010, Khodor *et al.* 2012, Tilgner *et al.* 2012]. However, introns annotated as alternative tend to be spliced post-transcriptionally, rather than co-transcriptionally [Vargas *et al.* 2011, Khodor *et al.* 2011, 2012, Tilgner *et al.* 2012]. How the spliceosome recognizes alternative exons and decides which exons and splice sites to include remains not fully understood [de Klerk and 't Hoen 2015].

1.2 NEXT-GENERATION SEQUENCING

Deep sequencing of cellular RNAs and genomic DNA has been widely used to study all aspects of gene expression [Reuter *et al.* 2015], including transcription and pre-mRNA splicing. The first steps towards this success were taken in the 1970s, when Frederick Sanger adopted the primer-extension strategy of replicating DNA with DNA polymerase and radio-labeled nucleotides to identify the underlying sequence. He included chain-terminating nucleotides (dideoxynucleotides) and made it possible to read out the DNA sequence by separating the nested PCR products in polyacrylamide gel electrophoresis [Sanger *et al.* 1977]. The replacement of radioactive with fluorescent labeling enabled the automation of the Sanger sequencing method and the development of “first generation sequencers”. Shortly afterwards, the so-called shotgun sequencing technique was developed, in which DNA is fragmented into smaller pieces, sequenced and then realigned in the downstream analysis [Anderson 1981]. Further developments lead to the first sequence of the human genome [Lander *et al.* 2001]. However, the method required molecular cloning of fragmented DNA pieces into bacterial artificial chromosomes, which was laborious and cost-intensive.

The limitations present in “first generation sequencing” have been overcome

with the development of “second generation sequencing” (or next generation/-high throughput) methods. For example, *in vitro* clonal amplification by bridge PCR of adaptor ligated DNA fragments on the surface of a glass slide was developed and then patented by Illumina¹. The clonal amplification results in the formation of local “DNA clusters”, whose underlying DNA sequence can be read in cycles of base incorporation, washing, imaging, and cleavage (sequencing by synthesis, Figure 2A). First, the reaction cell (flow cell) is flushed with reversible terminator nucleotides, which can be ligated by DNA polymerase depending on the template sequence. Next, non-incorporated nucleotides are washed away and the fluorescent signal of incorporated nucleotides is recorded. In the final step, fluorescent dye and terminal 3’ blocker are removed and the cycle starts anew. Despite the development of other sequencing techniques, Illumina remains the most common one (reviewed in [Mardis 2008, Reuter et al. 2015]).

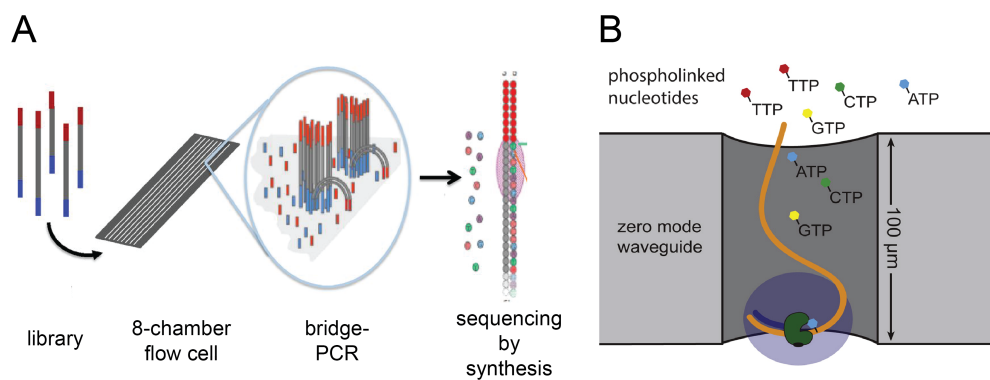


FIGURE 2 Next generation sequencing technology. A: High-throughput sequencing (Illumina technology). (c)DNA libraries carrying specific 5’ and 3’ adaptors are introduced into a microscope slide with 8 flow channels containing complementary oligonucleotides on their surfaces, which can anneal to the library ends. Bridge amplification is performed to amplify individual fragments of the library to enhance sequencing detection. On the sequencing machine, sequencing by synthesis takes place with n cycles of reversible terminator base incorporation, washing, imaging, and cleavage of the 3’ end group (n =read length). Image adapted from [Johnsen et al. 2013] B: Single molecule real time sequencing (Pacific Bioscience’s SMRT sequencing). A single DNA polymerase molecule is positioned at the bottom of a ZMW (zero mode waveguide). Phosphate-labeled versions of all four nucleotides are present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW and results in a fluorescent signal that is recorded with a camera. Image adapted from [Reuter et al. 2015]

A typical sequencing run with the Illumina HiSeq2500 technology² produces 1 Tb in 6 days [Reuter et al. 2015]. Thanks to the relatively low average error rate of less than 1%, the obtained short sequences can be aligned back to the genome using appropriate software [Dohm et al. 2008, Ruffalo et al. 2011]. Initially, high

1 www.illumina.com Kawashima, Eric H.; Laurent Farinelli; Pascal Mayer (2005-05-12). “Patent: Method of nucleic acid amplification”

2 www.illumina.com/systems/hiseq_2500

error rates towards higher cycle numbers limited the maximum read lengths to 25 bp. Technical improvements made it not only possible to acquire up to 150 bp reads, but also to sequence both ends of a DNA fragment bound to the flow cell (paired-end sequencing) [Kircher et al. 2011]. In order to analyze transcriptomes by short read sequencing, so-called RNA-Seq, specific library protocols have been developed. These protocols are often strand-specific and include fragmentation of RNA, reverse transcription into cDNA, specific adaptor ligation and cDNA enrichment by PCR [Levin et al. 2010].

Increasing read length to facilitate genome sequencing and assembly and alternative isoform detection in transcripts is still a focus of sequencing technology developments and other platforms exceed Illumina sequencing in this regard [Stranneheim and Lundeberg 2012]. One example is Single Molecule Real Time sequencing (SMRT sequencing) commercialized by Pacific Biosciences sequencing (PacBio), which allows to sequence DNA molecules up to several kilobases in length without prior amplification, albeit with less throughput [Eid et al. 2009, Korlach et al. 2010]. This technology can for example be used for *de novo* genome sequencing of small bacterial genomes [Chin et al. 2011], direct detection of cytidine methylations in genomic DNA [Flusberg et al. 2010] and even to detect splice isoforms in single polyadenylated mRNA transcripts [Sharon et al. 2013]. The accompanying template preparation involves the ligation of single-stranded hairpin adaptors onto the ends of double-stranded DNA, generating a circular template for sequencing (SMRT-bell). With a strand-displacing polymerase, the DNA molecule can be sequenced multiple times, which increases accuracy [Travers et al. 2010]. The DNA polymerase is attached to the bottom of a zero-mode waveguide (ZMW) [Foquet et al. 2008], a small nanometer sized pore, and can bind the DNA template. Nucleotides, which carry a fluorescent label on their phosphate chain (phospholinked nucleotides) are flushed into the SMRT cell. Upon nucleotide binding to the DNA template and the DNA polymerase, the residence time of the nucleotide close to the detection surface at the bottom of the ZMW increases and fluorescence can be recorded. Incorporation of the nucleotide into the nascent DNA chain removes the fluorescent label and ligation and fluorescence detection of the next nucleotide can occur (Figure 2). Overall, PacBio sequencing allows to obtain very long continuous sequence reads of often several kilobases. The error rate of a single read is relatively high. The circularization of template DNA with hairpin adaptors allows that DNA polymerase passes the template DNA several times. A consensus sequence generated from this decreases the error rate, but also the possible read length to 0.25-2 kb [Quail et al. 2012].

1.3 RNA SEQUENCING TO STUDY CO-TRANSCRIPTIONAL SPLICING

RNA-Seq of mRNA can be used to annotate and characterize introns and alternative transcript isoforms [Blencowe et al. 2009]. This gives information on final pre-mRNA splicing outcomes, but does not distinguish analysis of pre-mRNA splicing in association with the RNA synthesis and the process of transcription.

Therefore, a basic requirement for global analyses of co-transcriptional splicing is the purification of nascent RNA (Figure 3A). Common strategies are either the purification of nascent RNAs attached to cellular chromatin [Tilgner et al. 2012, Oesterreich et al. 2010, Khodor et al. 2011, Bhatt et al. 2012, Khodor et al. 2012, Bhorjee and Pederson 1973, Wuarin and Schibler 1994] or metabolic labeling and purification utilizing the specific label [Windhager et al. 2012, Heyn et al. 2014, Miller et al. 2011].

Distiguishing features of nascent RNA are a high variability in length (from a few nucleotides to thousands), different 3' ends reflecting the progression of Pol II in the gene, and the lack of polyadenylation. Nascent RNA coverage profiles over gene bodies usually show a saw-tooth pattern of read density, revealing ongoing transcription and co-transcriptional splicing (Figure 3B). A 5' to 3' gradient in read density is not observed in corresponding mRNA-Seq experiments [Carrillo Oesterreich et al. 2010, Khodor et al. 2011, Zaghlool et al. 2013].

The use of nascent RNA and the development of deep sequencing strategies made it possible to quantify steady-state co-transcriptional splicing levels. Short read sequencing (Section 1.2) results in millions of reads originating from the nascent RNA pool, reflecting unspliced and spliced populations of transcripts. Various strategies are currently used to analyze those data. Even though the results of several different experimental approaches and analyses agree on the widespread nature of co-transcriptional splicing, some disagree - and it remains unclear to what extent the existing analysis strategies are comparable [Brugiolo et al. 2013].

Several split-read sensitive tools are available for sequencing read alignment to the genome or transcriptome [Kim et al. 2013, Dobin et al. 2013, Hoffmann et al. 2014, Wang et al. 2010, Ameer et al. 2010]. In addition, numerous software to identify and quantify various mRNA splice isoforms, e. g. MISO, spliceR, SplicingCompass and Cufflinks, has been developed [Trapnell et al. 2012, Katz et al. 2010, Aschoff et al. 2013, Vitting-Seerup et al. 2014].

The above mentioned decay of nascent RNA-Seq coverage towards the 3' end of genes might influence the quantification of isoforms, as exons close to the 3' end of genes are likely to be less represented in the data. The comparison of gene and intron coverage has been used before to estimate global co-transcriptional levels in cells [Khodor et al. 2011, Bhatt et al. 2012], but this does not permit analysis of splicing efficiencies for individual introns. This is critically important, because intron retention or slow splicing is usually not uniform among all the introns of the same transcript [Tilgner et al. 2012, Pulyakhina et al. 2015, Schwarze et al. 1999].

Additional aspects of the protein-coding gene architecture might influence splicing estimates when coverage of full exons or introns is used. For example, highly abundant snoRNAs can lead to high, splicing-unrelated coverage within introns that is unrelated to splicing, yielding falsely reduced splicing efficiencies. In addition, large differences in sequence composition and length of introns and exons might alter the sequence coverage per feature in a non-controllable fashion. Furthermore, the presence of many unfinished transcripts in the nascent RNA fraction enhances intronic signal and might underestimate splicing values. Tran-

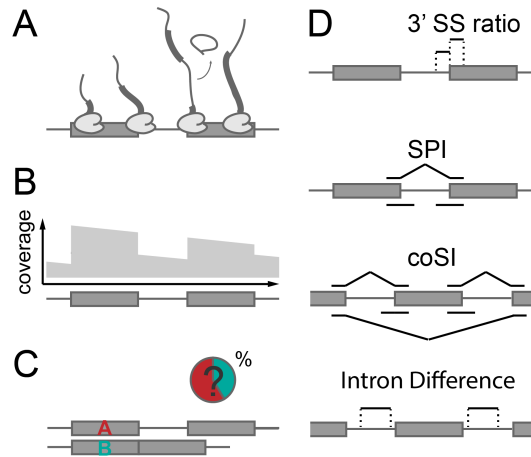


FIGURE 3 Global co-transcriptional splicing values from nascent RNA sequencing. A: Two exons of an intron-containing gene with RNA polymerase II molecules (Pol II, light grey) synthesizing nascent RNA. After the intron (thin line) is synthesized, it can be removed from the nascent RNA through splicing. B: Nascent RNA sequencing reads can be mapped to the gene body and differences in sequence coverage between introns and exons illustrate co-transcriptional splicing. 5' to 3' gradient of sequence coverage across the introns and exons reflects nascent transcripts in various stages of transcription. C: Sequence read counts of spliced and unspliced nascent RNAs are used to estimate co-transcriptional splicing. D: Four ways to estimate co-transcriptional splicing values per intron or exon, using either coverage per bp (3' SS ratio, Intron Difference) or spliced and unspliced junction read counts (splicing per intron [SPI], completed splicing index [coSI]).

scriptural pausing within a particular gene location is another uncontrollable variable that could increase or reduce the coverage in particular gene regions. In order to avoid these sources of error, alternative approaches to analyzing coverage over full introns or exons are intron- or exon-centric. Previously, four different strategies have been applied for the analysis of co-transcriptional splicing (Figure 3C-D, Chapter 9):

- A. Intron-centric 3' splice site ratio (3' SS ratio): The ratio between gene coverage in a 25 bp window upstream and downstream of the respective 3' splice site is calculated. This calculation strategy reduces the inclusion of intron signal, which results from elongating Pol II molecules that have not yet reached the 3' SS. Most introns and exons are longer than 25 bp, so that the window does not cover the next intron or exon [Khodor et al. 2011].
- B. Intron-centric splicing per intron (SPI): The number of spliced junction reads (covering the exon-exon junction) is compared to the total number of junction reads (covering the exon-exon and exon-intron junction). This measure can be calculated for each splice site individually or combined, whereas the unspliced fraction of reads must be divided by two to prevent double counting of unspliced events. Higher sequencing depth is required to use this junction-read based splicing metric compared to the 3' SS ratio metric. Each junction can be defined by a spliced read. In that way also

de novo intron detection and splicing calculation is possible [Grisdale et al. 2013].

- c. Exon-centric completed splicing index (coSI): Junction reads are also used to estimate splicing values, but for individual exons. Exon skipping events are included into the splicing measure. Splicing values for first and last exons cannot be considered [Tilgner et al. 2012].
- d. Exon-centric intron difference: To infer co-transcriptional splicing values from total RNA-Seq data, which include both nascent and mature messenger RNA, this approach uses the difference in sequence coverage between upstream and downstream intron. To estimate co-transcriptional splicing with this approach one assumes that the decrease in signal between a 500bp window upstream of the exon of interest to a similar sized window downstream of the exon is mainly due to co-transcriptional splicing and not just nascent transcription. This difference cannot be calculated for introns smaller than the window of analysis. Further it shows a distribution around 0 and is strongest for very long introns (> 50,000 nt) [Ameur et al. 2011].

1.4 TWO YEASTS AND THEIR GENE ARCHITECTURE

S. cerevisiae, as a single cell eukaryote, serves as a strong model system to study conserved cellular processes, e.g. the nature and dynamics of the spliceosomal network [Forsburg 2005, Wahl et al. 2009]. However, intron-containing genes are rare in *S. cerevisiae* and might represent a very specific class of genes with distinct properties. They are likely remaining genes that have not lost their introns upon homologous recombination from reverse transcribed cDNAs [Fink 1987]. Here are parallels to an evolutionary co-elimination of spliceosome components compared to fungal and non-fungal ancestors [Aravind et al. 2000]. Therefore it is necessary to combine findings from *S. cerevisiae* with the study of pre-mRNA splicing in other organisms, e.g. *S. pombe*, to grasp the full spectrum of pre-mRNA splicing and its integration into global gene expression. Already in other research aspects of cellular biology, e.g. cell cycle check point regulation, significant progress was made by comparing the distantly related yeasts *S. cerevisiae* and *S. pombe* [Forsburg 2005]. The two species separated in evolution about 1.14 Bya and their ancestor diverged from later metazoans about 1.52 Bya [Sipiczki 2000] (Figure 4A).

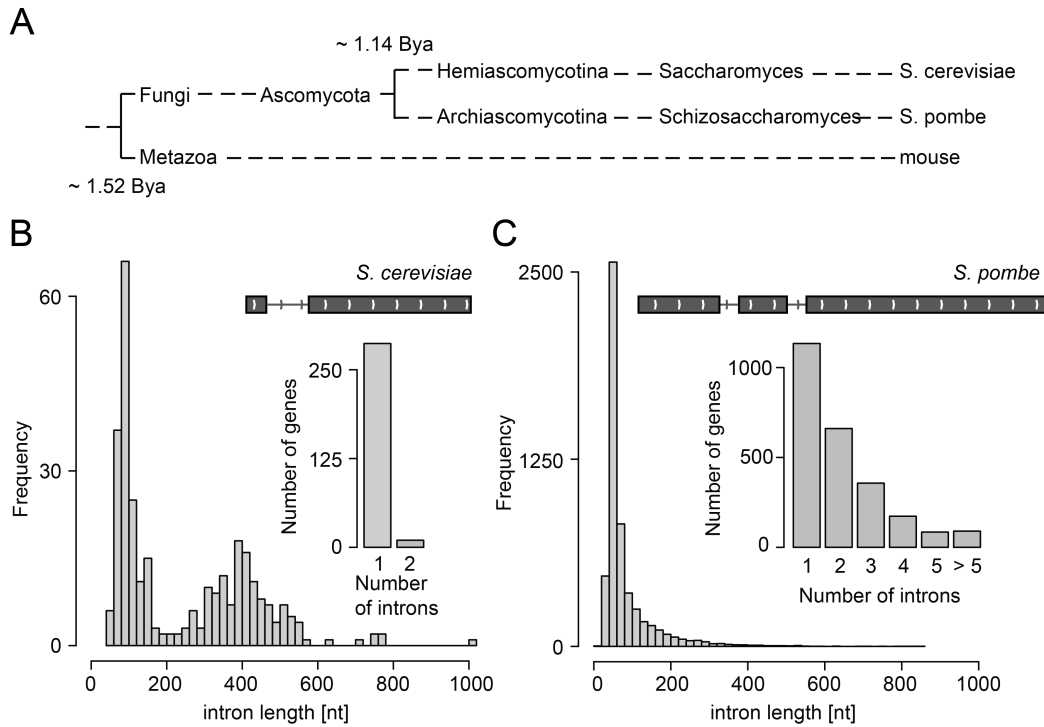


FIGURE 4 *S. cerevisiae* and *S. pombe* distance in evolution and gene architecture. A: *S. cerevisiae* and *S. pombe* are only distantly related and diverged in evolution ~1.14 billion years ago (Bya). Separation to metazoans occurred ~1.52 Bya (Figure after [Hedges 2002, Sipiczki 2000]). B: 296 intron-containing protein-coding genes are annotated (Scer3, 10 with 2 introns) and the intron length distribution follows a bimodal distribution (histogram). The median intron-containing *S. cerevisiae* gene is shown as a diagram. The total gene length (median 1,006 nt) is shorter than intron-containing genes in *S. pombe* (median 1,836 nt, see C). C: More than 5,300 introns are annotated in *S. pombe* (Ensembl annotation EF2). They are usually very short (median 56 nt) and half of the intron-containing genes contain two or more introns. The median *S. pombe* gene is longer than the median intron-containing *S. cerevisiae* gene (compare B).



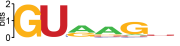



	S. CEREVISIAE	S. POMBE	M. MUSCULUS
genome size	12.1 Mb	12.5 Mb	2.7 Gb
number of chromosomes	16	3	20
number of genes	~6,000	~5,000	~25,000
number of introns	296	~5,300	>160,000
intron number/ gene	1	2	8
intron length	128 bp	56 bp	1,387 bp
first exon length	72 bp	232 bp	197 bp
internal exon length	101 bp	137 bp	126 bp
terminal exon length	582 bp	685 bp	939 bp
5' splice site			
3' splice site			

TABLE 1 Genome and gene features of *S. cerevisiae* (Scer3), *S. pombe* (EF2), *M. musculus* (mm9 RefSeq). The length values refer to medians. 5' and 3' splice site motifs were generated from all *S. cerevisiae* and *S. pombe* introns and 10,000 randomly chosen intron from *M. musculus* with Weblogo Version 2.8.2 (2005-09-08).

Currently 146 genes are annotated with the GO term "RNA splicing" in *S. pombe* (PomBase GO:0008380). Higher similarity to the human splicing factors than *S. cerevisiae* lead to the conclusion that *S. pombe* contains a splicing machinery that more closely reflects the archetype of a spliceosome machinery [Kaeufer and Potashkin 2000]. Further similarities between the gene architecture of *S. pombe* and metazoans are the number of intron-containing genes (~50%), the splice site conservation and the number and length of exons (Figure 4, Table 1). However, the short length of *S. pombe* introns is much more similar to other fungi species, rather than metazoans. It also differs from *S. cerevisiae*, which is an exception among fungi with a bimodal distribution of intron length ([Kupfer et al. 2004], Figure 4). Table 1 and Figure 4 highlight important similarities in gene architecture between three distantly related species, the two yeasts, *S. cerevisiae* and *S. pombe*, and mouse, which represents the mammalian gene architecture. Splice sites are very strongly conserved in *S. cerevisiae*, less in *S. pombe*, and in *M. musculus* only the first two nucleotides in 5' SS and 3' SS are strongly conserved in splice sites. The same trend is seen for the branchpoint motif in all three species [Kuhn and Käufer 2003].

THESIS AIM AND OVERVIEW

The process of pre-mRNA splicing has already been the focus of many studies. Yet, it remains unclear how and to which extent pre-mRNA splicing and transcription are functionally coupled, when it occurs during transcription and how alternative splicing can arise. In this thesis, I address the following questions: How prevalent is co-transcriptional splicing in *S. pombe* and what determines co-transcriptional splicing in this species? How does pre-mRNA splicing contribute to shape gene expression profiles? In which order are introns removed in an eukaryotic organism without prevalent alternative splicing? When during transcription does pre-mRNA splicing occur in endogenous genes in *S. cerevisiae* and *S. pombe*? In order to answer those outstanding questions, I decided to apply deep sequencing of nascent RNA. This will determine co-transcriptional splicing profiles with high resolution and accuracy.

By developing a new strategy to quantify co-transcriptional splicing, which is called Single Molecule Intron Tracking (SMIT), I aim to counteract the limitations of previous studies and gain unprecedented insights into this fascinating process. The following improvements over existing strategies will be crucial:

1. The (un)spliced transcript is directly identified by sequencing and not indirectly by linking it to a potentially interfering fluorescent label.
2. A diverse pool of transcripts originating from millions of unsynchronized cells at different stages of gene expression can be assayed, without artificial transcription repression or induction.
3. Co-transcriptional splicing of endogenous genes with different properties can be assayed.
4. The position of Pol II during transcription with the nascent RNA attached in the elongation complex can be identified with single nucleotide resolution.

Co-transcriptional splicing levels have been well characterized in *S. cerevisiae*, the gene architecture is simple with mainly single intron genes and also many other aspects of transcription and chromatin biology are well understood in this model organism [Botstein and Fink 2011], which can help later on to place the kinetic data on co-transcriptional splicing into context. Therefore, *S. cerevisiae* forms an ideal organism to establish such a novel approach of using deep sequencing to obtain kinetic information on a cellular process. I extend co-transcriptional splicing analysis to the fission yeast *S. pombe*, a single-cell eukaryote with many multi-intron genes and a gene architecture more similar to higher eukaryotes.

The results part of this thesis is divided into three chapters:

[Chapter 3](#) evaluates existing approaches (introduced in [Section 1.2](#)) to quantify co-transcriptional splicing from nascent RNA-Seq experiments. This is done on two previously published mouse nascent RNA-Seq datasets. The comparison assesses to which extent the different published co-transcriptional splicing calculation strategies impact outcome and can be used to quantify co-transcriptional in *S. pombe* and *S. cerevisiae*. The results of this section have been published recently [[Herzel and Neugebauer 2015](#)]. The second part of the chapter focuses on the establishment and characterization of a cell fractionation protocol in *S. pombe* and nascent RNA preparation from *S. pombe* chromatin. Subsequently, I present the nascent and mRNA *S. pombe* sequencing data and the quantification of pre-mRNA splicing in *S. pombe*.

[Chapter 4](#) addresses the question, what determines co-transcriptional splicing in *S. pombe*. The nascent and mRNA-Seq data from [Section 3.2](#) are analyzed with regard to gene function, expression and gene architecture. The second part addresses how co-transcriptional splicing and other RNA processing events are linked. In particular, circular RNA formation, the order of intron removal in multi-intronic transcripts and nascent RNA cleavage at the polyA site are considered. To analyze two RNA processing events on the same transcript a novel strategy of sequencing nascent RNA is developed.

[Chapter 5](#) focuses on the development of a method to quantify the progression of co-transcriptional splicing with respect to the distance to the intron end. Then I present data on individual genes and highlight common and distinctive properties in co-transcriptional splicing patterns of all assayed genes in *S. cerevisiae*.

The discussion part is divided into two chapters:

In [Chapter 6](#) I evaluate my results and findings on the prevalence and position of co-transcriptional splicing during transcription in *S. cerevisiae* and *S. pombe* with respect to previous knowledge on transcription elongation and PolII pausing. The analysis of *S. pombe* co-transcriptional splicing is discussed with respect to findings in other species and which aspects reveal conserved principles or species-specific aspects of co-transcriptional splicing.

In [Chapter 7](#) I will discuss properties and future potential of the methods, which I developed as part of this work to address questions on timing and association of co-transcriptional with transcription and other RNA processing events.

I will summarize and end this thesis with concluding remarks ([Chapter 8](#)). The remaining parts of the thesis contain material and methods ([Chapter 9](#)) and supplementary information for each results chapter ([Appendix A](#), [Appendix B](#), [Appendix C](#)) and oligonucleotides used in the experiments presented in this thesis ([Appendix D](#)).

Part II
RESULTS

QUANTIFICATION OF CO-TRANSCRIPTIONAL SPLICING

3.1 STRATEGIES TO QUANTIFY CO-TRANSCRIPTIONAL SPLICING

Recently, a number of studies using global methods have shown that the majority of splicing is co-transcriptional, yet not all published studies agree in their conclusions and it is unclear how the chosen ways to quantify co-transcriptional splicing compare to each other. Short read sequencing of RNA (RNA-Seq) is the prevailing approach to measuring splicing levels in nascent RNA, mRNA or total RNA. Here, I compare four different strategies for analyzing and quantifying co-transcriptional splicing.

3.1.1 *Nascent RNA-Seq in previous studies*

I have chosen two published nascent RNA-Seq datasets for comparative analysis [Bhatt et al. 2012, Khodor et al. 2012]. In both studies, chromatin-associated RNA was extracted from mice and compared to polyA+ RNA samples from either total RNA or cytoplasmic RNA. However, one was done in liver tissue [Khodor et al. 2012] and one in untreated macrophage cells [Bhatt et al. 2012].

In the chosen datasets, I quantified splicing levels for ~100,000 (55% of all annotated, non-redundant) introns from 140 Mio 80-100bp reads, using a coverage-based splicing measure [Khodor et al. 2012], but only ~34,000 (18% of all annotated, non-redundant) introns from ~43 Mio 50bp reads [Bhatt et al. 2012].

3.1.2 *Quantification of co-transcriptional pre-mRNA splicing in previous studies*

Both datasets were mapped and analyzed in the same way to ensure comparability. Tophat2 [Kim et al. 2013] was used for mapping with stringent settings, e.g. no mismatches and no multimapping, and a gene model annotation¹ was provided to ensure mapping to known transcript models first before raw data are mapped to the genome and novel junctions. All four splicing calculation procedures were applied. One of the four approaches, the 3' SS ratio, was developed to estimate the unspliced fraction of introns (Section 1.3). Here, I subtract those values from 1, yielding an estimate of the fraction of completed splicing and distinguish this from completely unspliced introns (3' SS ratio < 0). Now it is possible to compare the 3' SS ratio to other approaches that calculate the fraction of completed splicing, rather than the unspliced fraction. The results are shown as cumulative distributions in Figure 5. Steady-state RNA splicing in samples containing polyA+ RNA, total (Figure 5A-B), cytoplasmic or nucleoplasmic (Figure 5D-E) mRNA is mostly complete with a median of 1. The distribution

¹ mm9 RefSeq: <https://genome.ucsc.edu/>

median for nascent RNA is significantly lower ranging from 0.57 to 0.63, indicating that 50% of introns or exons are spliced to 60% co-transcriptionally. However, the median splicing value (0.90) for chromatin-associated RNA from mouse macrophages is substantially higher (Figure 5D-E).

[Khodor et al. 2012] noted a 5' to 3' abundance gradient in nascent RNA-Seq gene coverage. This gradient is especially pronounced in very long introns and can be used to detect co-transcriptional splicing of adjacent exons similar to [Ameur et al. 2011]. Figure 5C displays data for exons with sufficient sequence coverage and downstream introns of 10,000 nt or longer. Splicing estimates for only a few exons for the nascent RNA sample ($n = 3,333$) and even fewer ($n = 557$) exons in mRNA can be calculated. Nevertheless, the median difference between intron signal around a certain exon is significantly higher for the nascent RNA sample (0.009) than for mRNA (0.004) (Figure 5C) indicating co-transcriptional splicing also with this limited calculation strategy. This is also true for the mouse macrophage dataset (chromatin-associated RNA (P1) 0.006 and cytoplasmic mRNA (C1) 0.002, Figure 5C). Differences between the two mouse samples could be of technical or biological nature (discussed in Section 7.1). However within each sample, cumulative distributions and median are very similar between the two intron-centric approaches (3' SS ratio and SPI) and the exon-centric approach (coSI). CoSI measures the fraction of completed splicing of two introns surrounding one exon, the 3' SS ratio takes read coverage information in a short window in intron and exon to calculate the fraction of intron splicing and the SPI (splicing per intron) utilizes junction reads (introduced in Section 1.2).

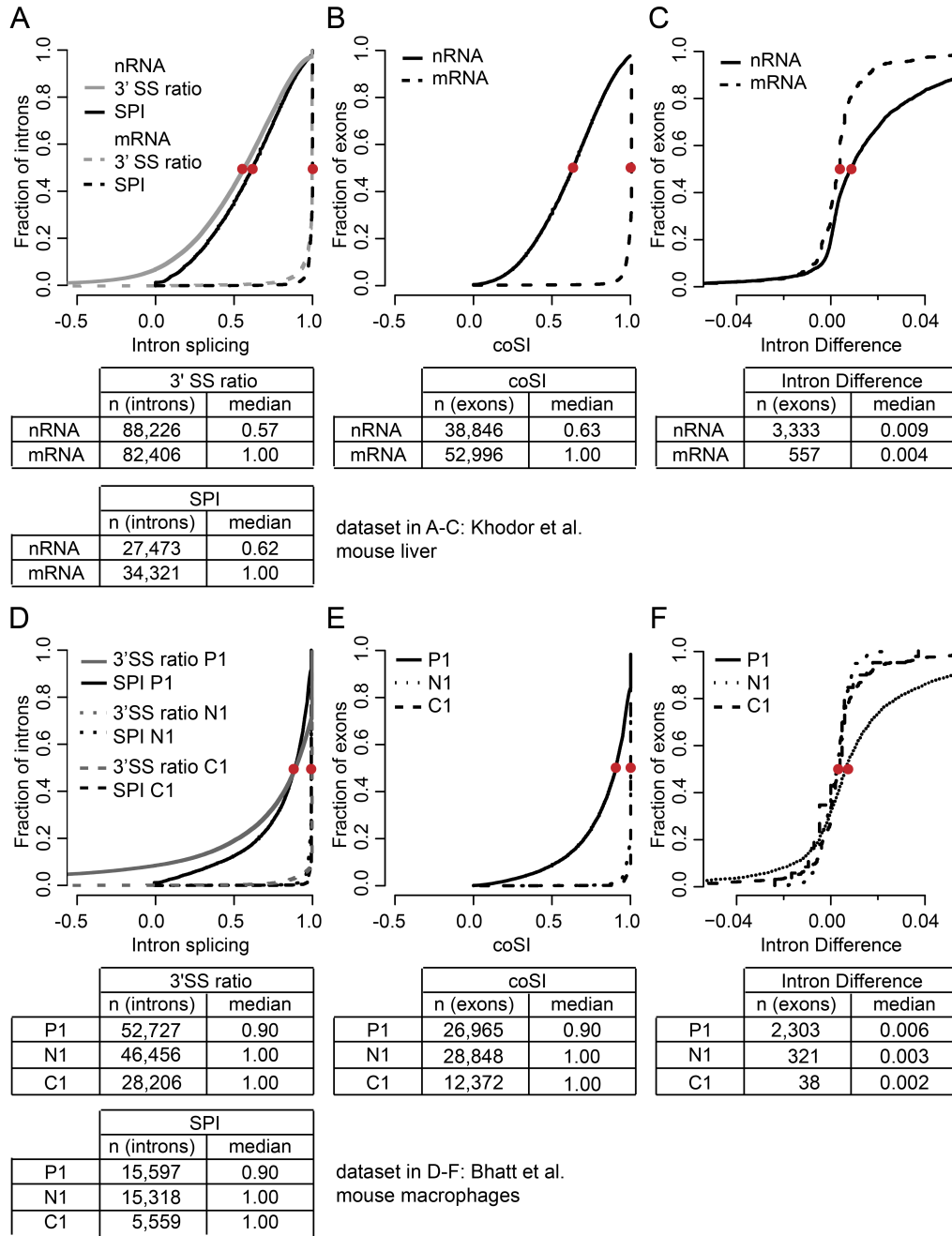


FIGURE 5 Global co-transcriptional splicing values in mouse liver and macrophages calculated with four different approaches. A,D: Cumulative distributions of intron-centric 3' SS ratio and SPI have very similar median co-transcriptional splicing values. Total mRNA splicing is complete with a median of 1.0. The number of introns included in the two analyses differs strongly. B,E: coSI distributions for nascent and mRNA samples have a similar median like the intron-centric splicing measures under A. C,F: The difference in intron coverage shows co-transcriptional splicing, but the number of exons included in the analysis differs widely from the other approaches. Only downstream introns equal or greater to 10 kb are included. A-C: data analyzed from [Khodor et al. 2012]. D-F: data analyzed from [Bhatt et al. 2012], P1 - chromatin, N1 - nucleoplasm, C1 - cytoplasm. Red dots indicate the median in co-transcriptional splicing values.

3.1.3 Transcription-associated aspects of quantification

Depending on intron size, the 5' to 3' abundance gradient (due to on-going transcription) might overestimate the amount of unspliced transcripts around the 5' SS. This is true for introns of several thousand nucleotides included in the nascent RNA/chromatin-associated RNA analysis, but is not seen in mRNA samples (Figure 6). To derive this conclusion, I calculated the 3' SS ratio per intron and, in a similar way, the respective 5' SS ratio. The distribution of differences between those two ratios should be 0, if transcription does not influence the intron splicing calculation and centered around smaller values than 0, if the transcription has an impact on intron splicing calculation, because the unspliced fraction of reads increases around the 5' SS. The difference between 5' SS ratio and 3' SS ratio was determined and the data were split into 10 evenly sized groups according to the associated intron length. The difference between 5' SS ratio and 3' SS ratio for both mRNA samples is indeed tightly centered around 0 for all groups of intron length (Figure 6C-D). The distribution for the nascent RNA/chromatin-associated RNA samples is broader and for long introns a significant difference between 5' SS ratio and 3' SS ratio is observed (Figure 6A-B).

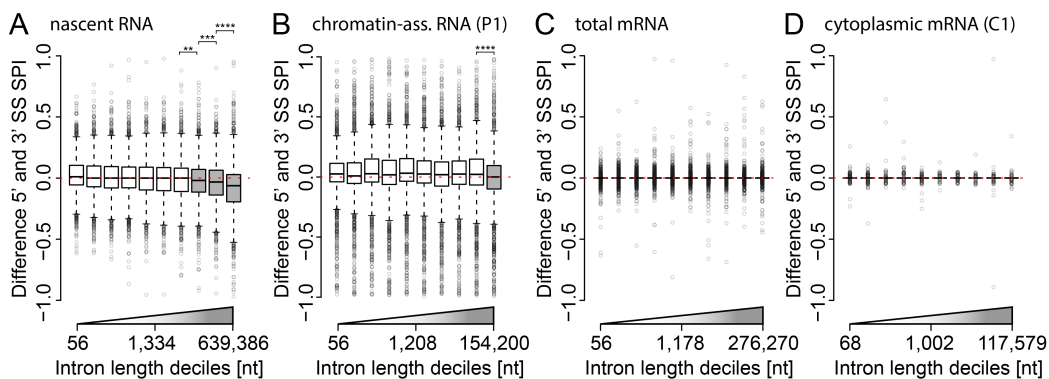


FIGURE 6 Splicing per intron (SPI) score is influenced by the 5' to 3' abundance gradient in long introns. Boxplots show the difference between 5' SS and 3' SS SPI for introns split into 10 evenly sized groups according to intron length. Boxwidth is proportional to the square root of the number of introns per group. Very and extremely significant differences between adjacent groups are indicated with asterisks ($p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****). Differences between groups have been tested with the Wilcoxon rank sum test. Red dotted line is plotted at 0 difference between both 5' and 3' SS SPIs. Grey boxes indicate long intron splicing values affected by 5' to 3' gradient. The minimum, median and maximum intron sizes given on the x-axis differ between different samples, because all four groups contained partially different introns due to differences in sequencing depth. A: Nascent RNA-Seq data from [Khodor et al. 2012], 27,473 constitutive non-redundant introns. B: Chromatin-associated RNA-Seq data from [Bhatt et al. 2012], 15,597 constitutive non-redundant introns. C: mRNA-Seq data from [Khodor et al. 2012], 34,321 constitutive non-redundant introns. D: Cytoplasmic mRNA-Seq data from [Bhatt et al. 2012], 5,559 constitutive non-redundant introns.

Therefore, this splicing score calculation is optimal for shorter introns in higher eukaryotes or in species containing only very short introns, e.g. fungi [Grisdale

et al. 2013]. In this general comparison between methods though I calculate splicing per intron including both the 5' and 3' splice site junction reads.

3.1.4 Comparison between splicing measures

One robustness test for the different splicing determinations is that the same data analyzed in a complementary way should yield similar co-transcriptional splicing efficiencies. For example, SPI and 3' SS ratio correlate well (Figure 7A), suggesting the obtained values represent the true splicing efficiency for each intron. However, coSI and the Intron-Difference values do not correlate well. This may be due to the presence of only very few exons in both calculations, and the two compared splicing measures have a very different range and window size. Better correlation is observed between coSI for different biological samples (P1 and nRNA) and for the Intron Difference correlation for the same two different chromatin-associated RNA samples even though they originate from different cell types, mouse macrophages and liver, and were generated by two different labs. Further, one of those samples is not depleted for mature polyA+ transcripts. [Ameur *et al. 2011*] used the Intron Difference to infer co-transcriptional splicing from total RNA samples, where the majority of transcripts are polyadenylated. Thus the correlation between liver nascent RNA and chromatin-associated RNA (Figure 7B) from macrophages suggests that co-transcriptional splicing is quite similar for the exons detected and correlated in both datasets. A full analysis of tissue and cell specific differences between co-transcriptional splicing in both systems cannot be achieved with these two datasets, due to strong differences in RNA preparation and deep sequencing.

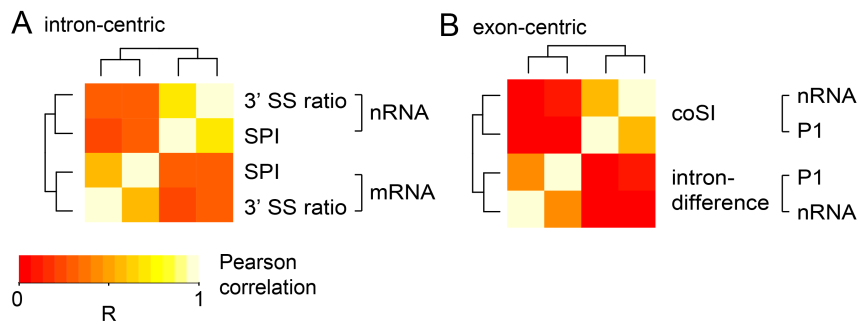


FIGURE 7 Correlation between different splicing measures. A: Intron-centric methods correlate well independent of the RNA sample, $n = 32,983$. B: Exon-centric approaches do not correlate among each other but rather between biological samples from different studies (P1 - chromatin-associated RNA from [Bhatt *et al. 2012*] and nRNA - nascent RNA from [Khodor *et al. 2012*]), $n = 302$. Heatmaps are derived from Pearson correlations.

3.1.5 *Applicability to yeast pre-mRNA splicing*

Overall, it becomes apparent that all strategies have certain limitations and advantages depending on the composition of the RNA samples and the biological question. The comparison of four possible ways to determine co-transcriptional splicing from nascent RNA-Seq data leads me to favor intron-centric approaches for the following reasons: First, intron-centric approaches make it possible to observe that co-transcriptional splicing varies not only between genes, but also between individual introns within one gene [Khodor et al. 2011]. Exon-centric approaches do not allow conclusions about splicing of individual introns and first and last exons, because splicing catalysis of two introns is combined within one splice score. Indeed in my analysis 3' splice site SPI of a downstream intron and 5' splice site SPI of the upstream intron do not correlate well (Section A.2, Figure 35), thus coSI and the Intron Difference estimate the chance of an exon to be associated with spliced introns, but averages splicing outcome of individual introns. In higher organisms intron boundaries are usually defined by exon definition, where the 3' splice site of an upstream intron is detected in connection to a proceeding 5' splice site. This links the process of splicing between two adjacent introns and forms an argument to calculate splicing with an exon-centric measure.

In order to analyze co-transcriptional pre-mRNA splicing in *S. pombe* or *S. cerevisiae* in the following parts of my thesis, the intron-centric SPI is appropriate and will be used. *S. pombe* and *S. cerevisiae* pre-mRNA splicing is assumed to be defined primarily through intron definition. Length dependencies of the different splicing measures do not apply to the two yeast species, as they harbor very short introns compared to mammalian protein-coding genes. Thus it is appropriate to use the combined SPI of 5' SS and 3' SS to derive splicing measures.

3.2 QUANTIFICATION OF CO-TRANSCRIPTIONAL SPLICING IN *s. pombe*

In mouse, fly, human and yeast co-transcriptional splicing, it was seen that the majority of introns are spliced co-transcriptionally with median co-transcriptional splicing value ranging from 60% to 75% [Brugiolo et al. 2013]. It remains unclear to which extent are introns removed co-transcriptionally in *S. pombe*, a promising model organism to study splicing of multi-intronic genes in absence of alternative splicing.

3.2.1 *Preparation of nascent RNA from S. pombe chromatin*

The first step along the way to analyze co-transcriptional splicing levels is to isolate nascent RNAs, which are attached to the gene locus via Pol II and are a class of very rare and transient of cellular RNA. In *S. cerevisiae* a cell fractionation scheme was developed previously to enrich for this RNA species [Carrillo Oesterreich et al. 2010]. I adapted the protocol for purification of nascent RNA from *S. pombe* chromatin. Data characterizing each step are shown in Figure 8. In particular, growth conditions and lysis (Figure 8B-C) had to be changed. Purification of

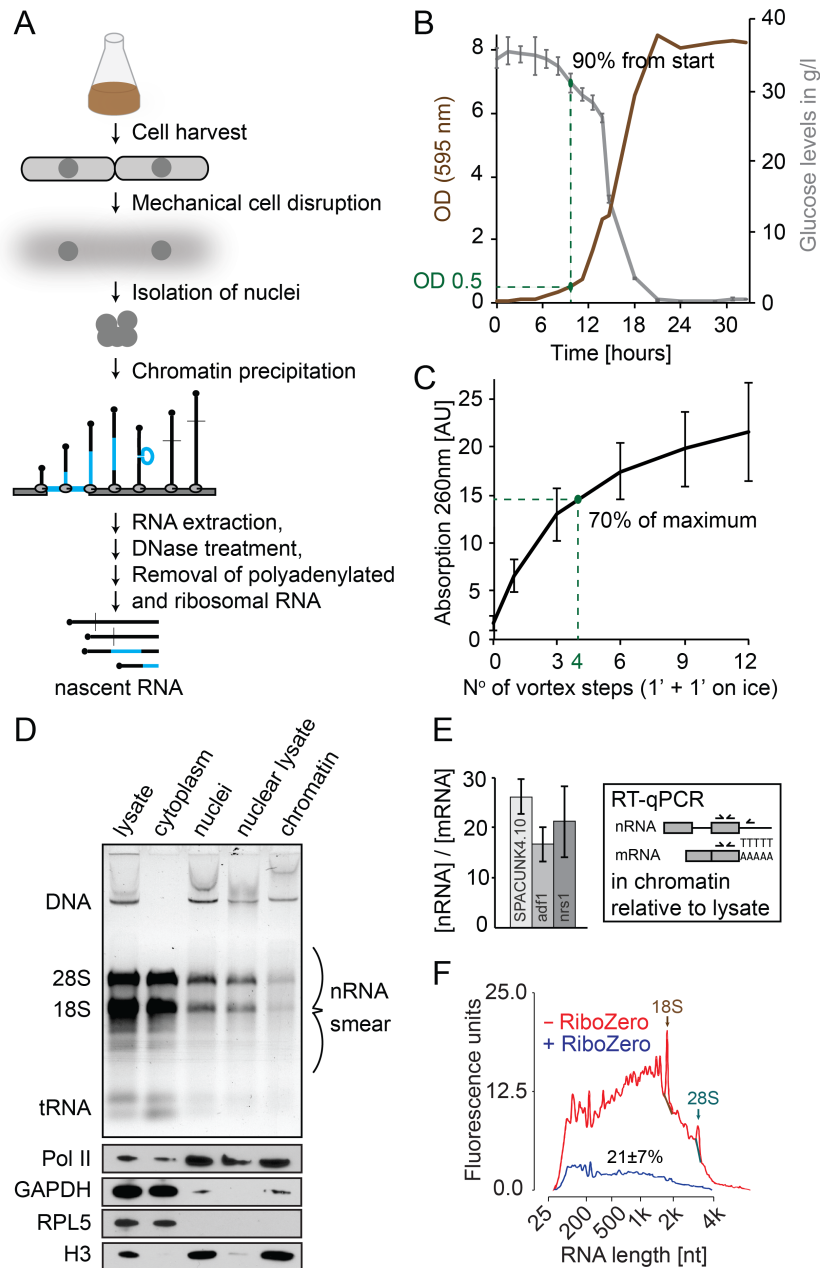


FIGURE 8 Biochemical enrichment of nascent RNA from *S. pombe* cells. **A:** Nascent RNA purification scheme starting from exponentially growing cells. **B:** *S. pombe* growth curve in YES and associated Glucose consumption profile (SD of technical replicates (n=3) shown). *S. pombe* cells are generally harvested at OD 0.5 after at least one cell number duplication. **C:** Nucleotide content of cell lysate after different vortex intervals. After 4 vortex intervals ~70% of cells are lysed compared to lysis in 16 vortex intervals (SD is shown, n=5). **D:** Nascent RNA purification from the chromatin fraction. Nucleotide analysis to assess enrichment of genomic DNA (DNA) and nascent RNA and the depletion mature rRNA (18S & 28S) and tRNA. Western blot analysis shows the enrichment of chromatin-associated proteins Pol II and Histone 3 (H3) and the depletion of cytoplasmic marker proteins GAPDH and RPL5. **E:** RT-qPCR to measure nascent RNA enrichment over mRNA for 3 genes after polyA+ mRNA depletion of the chromatin fraction. RT primers downstream of the polyA cleavage sites target nascent RNA, and oligo(dT) RT primer targets mRNA, respectively (SEM is shown, n=3-6). **F:** Mature rRNA and its precursors is efficiently removed following the RiboZero protocol. Characteristic peaks in the pattern of nascent RNA are kept throughout the depletion (SD is shown, n=7).

chromatin was validated using agarose gel electrophoresis, western blot analysis (Figure 8D-E) and mass spectrometry.

Mass spectrometry analysis of 2 chromatin samples and 1 cytoplasm sample as reference identified 421 proteins more abundant in chromatin than cytoplasm (Figure 9, Table 2). I required that those protein hits were detected by at least two unique peptides and that they have a chromatin/cytoplasm ratio of peptide counts greater than one. The GO term analysis confirms the specificity of the preparation, with chromatin-, nucleolus- and splicing-associated terms being the most significant hits for the chromatin fraction and cytoplasm-, proteasome- and mitochondria- and vesicle trafficking-associated terms being the most significant hits for the cytoplasmic fraction (Figure 9).

The comparison to *S. cerevisiae* chromatin data [Carrillo Oesterreich et al. 2010] shows that a very similar number of proteins is identified (425 in *S. cerevisiae* and 421 in *S. pombe*) and also similar enrichment for components of nucleosomes, DNA replication machinery, RNA polymerases, chromatin remodelers and ribosome biogenesis factors is observed (Table 2). However, the MCM complex involved in DNA replication is not identified in *S. pombe*, but is present with all its subunits in the *S. cerevisiae* chromatin fraction. The opposite is true for spliceosomal snRNP components, which are completely absent in *S. cerevisiae*, but enriched in both *S. pombe* chromatin replicates.

From the chromatin fraction, RNA and DNA are extracted and genomic DNA is digested in two subsequent DNase treatments with RNA column purification inbetween. Terminated and polyadenylated transcripts, which are present in the complex pool of chromatin-associated RNA, are removed in three sequential oligo(dT)-selection steps. I assessed enrichment of full-length non-terminated nascent RNA over polyadenylated RNA with a RT-qPCR assay, in which RT primers are designed to anneal downstream of the poly(A) cleavage sites to target nascent RNA and polyadenylated RNA is reverse transcribed using an oligo(dT) RT primers. A 17 to 30-fold enrichment depending on the assayed gene and RT-primer concentration is observed (Figure 8E).

In contrast to the microarray technology deep sequencing applications reveal potentially every sequence present in the sample. This has great advantages for *de novo* identification of transcripts, but also requires removal of very abundant sequences, e. g. mature rRNA and its precursors, which can prevent detection of less abundant RNAs, e. g. protein-coding nascent RNAs. Therefore, rRNA removal is included in the nascent RNA preparation (Figure 8F) and depletion is assessed using for example the Bioanalyzer platform.

With this preparation on hand, RNA samples were prepared for different sub-cellular fractions, cytoplasm, nucleus and chromatin, and submitted for strand-specific RNA-Sequencing.

3.2.2 *S. pombe* nascent and mRNA-Seq

On average 14 Mio 76 nt single-end reads were generated per sample (Section B.1). Each experiment was sequenced in three biological replicates. Transcriptome mapping was carried out with Tophat2 [Kim et al. 2013] with 2 general mis-

		<i>S. cerevisiae</i>	<i>S. pombe</i>
Nucleosomes	Histones	6/6	4/6
Replication machinery	Pol alpha	-	-
	Pol delta	-	-
	Pol epsilon	1/5	-
	MCM complex	6/6	-
	DNA replication factor C complex	5/6	5/6
	DNA replication factor A complex	2/3	1/3
RNA polymerases	Pol I	12/14	10/14
	Pol II	10/12	8/12
	Pol III	12/15	7/18
Splicing factors	U1	-	3/9
	U2	-	6/17
	Prp19 complex	-	9/14
	U5	-	2/6
	Sm-Ring	-	6/7
	SR-like	-	1/2
Chromatin remodelers	RSC	14/14	12/13
	yINO80	10/14	11/14
	SWI/SNF	8/12	7/12
	ISW1	4/4	NA
	ISW2	3/4	NA
	CHD-type	1/1	2/2
	FACT	1/2	2/2
Ribosome biogenesis		163/429	144/315
other		167	189

TABLE 2 Mass spectrometry of *S. cerevisiae* and *S. pombe* chromatin. *S. cerevisiae* data taken from [Carrillo Oesterreich et al. 2010]. 421 proteins were found to be enriched in at least one of 2 *S. pombe* chromatin replicates compared to cytoplasm (2nd replicate also enriched or equal to cytoplasm). The majority of hits could be grouped according to chromatin-associated functions or components.

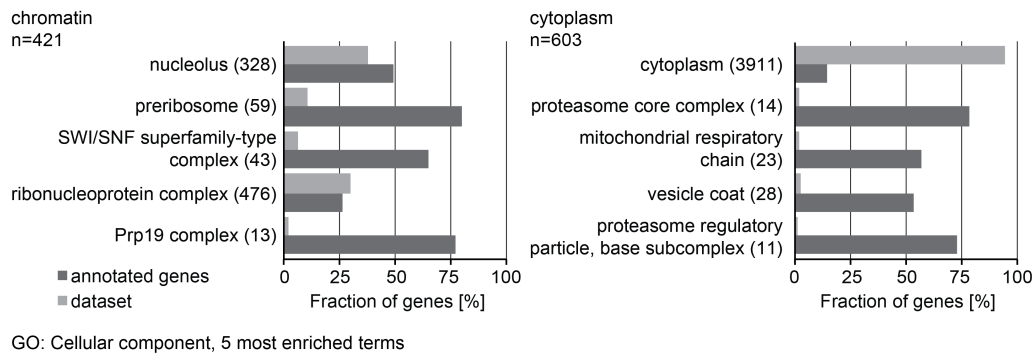


FIGURE 9 GO analysis of chromatin and cytoplasmic fraction. GO analysis was performed with topGO package in R. The five most enriched terms for “Cellular component” are shown. Number of genes annotated for each term are given in brackets.

matches, but no splicing mismatches allowed. Intron length settings were adjusted to *S. pombe* specific gene architecture with introns ranging from 30 nt to 900 nt and the genome annotation EF2 was provided to aid spliced transcript mapping. Mapping efficiency was greater than 95% (Table 10).

Normalized RNA-Seq coverage data for individual genes reveal characteristic patterns associated with the maturation state of protein-coding transcripts in the different cellular fractions (Figure 10A). Read coverage over the two introns shown in the example is lower compared to adjacent exons, suggesting pre-mRNA splicing in all three fractions. Intron coverage decreases from the chromatin-associated nascent RNA sample to nuclear mRNA and is not visible in cytoplasmic mRNA. Read coverage downstream of the annotated end of the gene is observed in chromatin-associated nascent RNA, but not in nuclear or cytoplasmic mRNA, and indicates the presence of long, not yet cleaved, non-polyadenylated transcripts in the nascent RNA fraction and/or on-going transcription after polyA site cleavage.

Figure 10B shows clustering of Pearson correlations of gene expression values calculated with Cufflinks and allows to assess the degree of correspondence between biological replicates. Overall the correlations between replicates and cellular fractions are high (> 0.6), indicating that gene transcription, which is assayed by profiling nascent RNAs, is a major determinant for steady state mRNA expression values. Biological replicates correlate highly among each other and less well with samples originating from different cellular fractions, indicating that sample quality is high and all replicates can be included in the further analysis. The single gene example from Figure 10A suggests a 5' to 3' coverage gradient in the nascent RNA sample, which could reflect the process of transcription. For mRNA this gradient should not be apparent. Figure 10C shows expression-normalized average nascent RNA and mRNA coverage profiles for the 50% highest expressed protein-coding intronless genes ranging from 1 to 4 kb. The annotated transcripts were scaled to a common start and end with 100 nt upstream and downstream of the annotated boundaries. This robustly shows the 5' to 3' coverage gradient for the nascent RNA fraction, but a 3' coverage bias in mRNA. Furthermore, higher read coverage at transcript boundaries in the

nascent RNA sample compared to mRNA is detected indicating the presence of non-terminated nascent transcripts. This and the coverage gradient nicely validate the purity of the nascent RNA pool compared to mRNA and allows to proceed with co- and post-transcriptional pre-mRNA splicing calculation for individual introns.

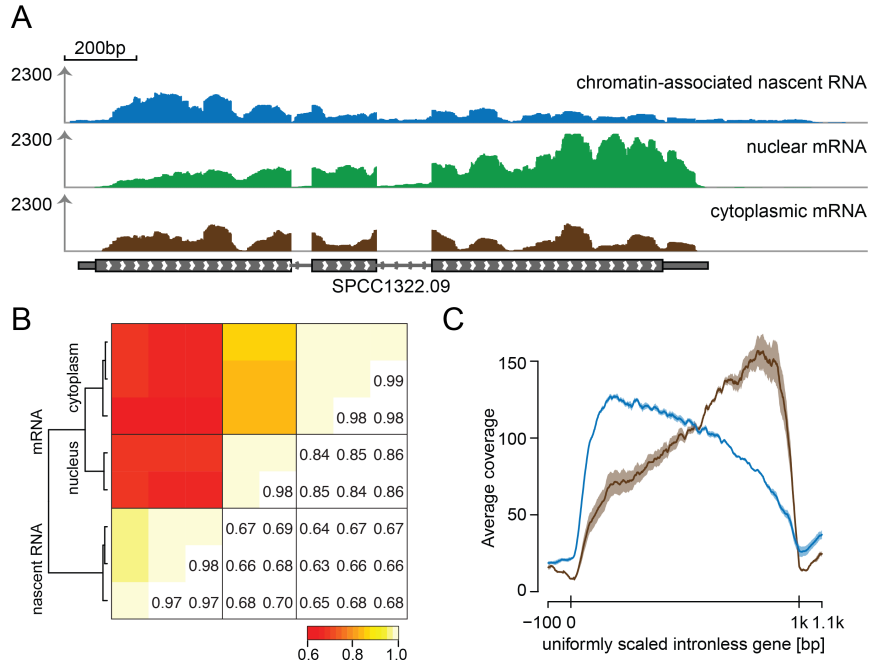


FIGURE 10 *S. pombe* RNA-Seq data from three cellular fractions. A: Read coverage over a two-intron gene for the three RNA samples. Counts per nucleotide are normalized to library size and coverage for one replicate of each cellular fraction is shown. B: Pearson correlation of gene expression values (calculated with Cufflinks) for all replicates included in the analysis. C: Average coverage profile over 668 high-expressed intronless genes ranging from 1-4 kb. Mean values with SD are shown ($n=3$). Nascent RNA coverage is shown in blue and mRNA coverage in brown.

3.2.3 Global pre-mRNA splicing estimates from chromatin-associated RNA and mRNA

Pre-mRNA splicing levels per intron were calculated in a similar way to what was referred to as “splicing per intron” in the previous section (Section 1.3). A low depth data cutoff of 10 junction reads was required. Figure 11A shows the cumulative distribution of SPIs for the three fractions. 5,300 introns are annotated in the gene annotation EF2 and the data allow to estimate intron splicing values for 43% (cytoplasm) to 90% (nascent RNA) of all introns with the requirement that each replicate contains sufficient number of reads to calculate the SPI. The degree of variation between replicates is very different for the different samples. The standard deviation of mRNA replicates follows an exponential distribution indicating very low variability between the different replicates. Nascent RNA replicates show a normal distribution of standard deviation with a mean of 0.1 reflecting higher degree of variation between replicates, which sets limitations

on the analysis of modest differences in co-transcriptional splicing. The median SPIs range from 0.58 (nascent RNA), 0.83 (nuclear mRNA) to 0.95 (cytoplasmic mRNA) showing that pre-mRNA splicing is co-transcriptional for the majority of introns. Most cytoplasmic mRNAs are fully spliced, whereas nuclear mRNA contains still a larger fraction of unspliced RNAs, which could still be spliced before nuclear export or could be degraded by RNA surveillance mechanisms.

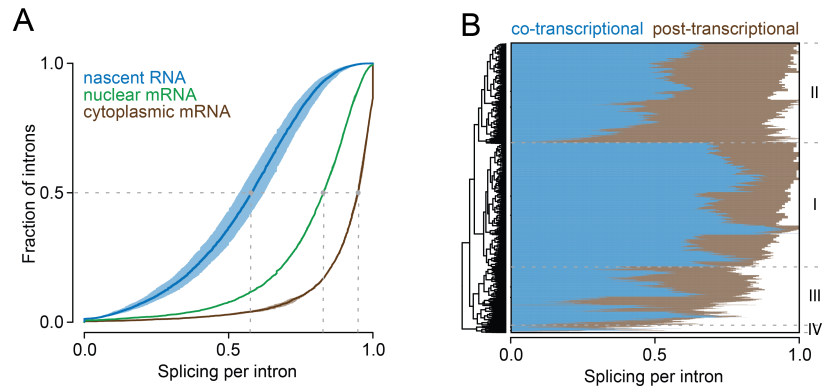


FIGURE 11 Splicing of *S. pombe* introns. A: Cumulative distribution of splicing per intron values for the majority of introns in *S. pombe* ($n(\text{nascent})=4,770$, $n(\text{nucleus})=4,339$, $n(\text{cytoplasm})=2,282$). Mean values with SD are shown ($n(\text{nascent}, \text{cytoplasm})=3$, $n(\text{nucleus})=2$). Grey dashed line and dots indicate the median SPIs for nascent RNA (0.58), nuclear mRNA (0.83) and cytoplasmic mRNA (0.95). B: Heatmap clustering and grouping of introns with respect to their fraction of average co- and post-transcriptional intron splicing ($n=3,165$). Groups were labeled according to their median levels of co-transcriptional splicing.

Figure 11B allows to assess to which extent individual introns are spliced co- and/or post-transcriptionally. Nascent RNA and cytoplasmic mRNA SPIs were hierarchically clustered. Four main clusters become apparent:

- I almost complete pre-mRNA splicing with large fraction of co-transcriptional splicing ($n=1,349$, 43% of quantified introns)
- II almost complete pre-mRNA splicing with large fraction of post-transcriptional splicing ($n=1,093$, 35% of quantified introns)
- III low co- and post-transcriptional splicing ($n=639$, 20% of quantified introns)
- IV retained introns ($\text{SPI} < 0.5$) ($n=84$, 3% of quantified introns)

60% of all annotated introns were subjected to the clustering analysis. Only 58 introns (0.01%) showed equal or higher co-transcriptional SPIs than the cytoplasmic SPIs and were excluded from clustering. Most introns (78%) fall into group I & II, which both show a very high amount of spliced transcripts in cytoplasm. This underlines that the majority of the intron-containing transcripts in *S. pombe* are spliced completely when they enter the cytoplasm and serve as template for protein synthesis. Groups I & II differ in the amount of co-transcriptional

splicing (median SPI 0.73 (group I) and 0.51 (group II)) and only very few introns are completely spliced co-transcriptionally (153 introns (5%) with >85% co-transcriptional splicing). Hence, one can conclude that pre-mRNA splicing for *S. pombe* introns starts co-transcriptionally and is completed post-transcriptionally in most cases.

3.2.4 *Patterns of co- and post-transcriptional splicing*

The nascent RNA preparation from *S. pombe* chromatin has been developed to estimate global co-transcriptional splicing levels and to study coupling between transcription and pre-mRNA splicing in this yeast species. The protein composition of the chromatin fraction is overall very similar to *S. cerevisiae* chromatin, but unlike to *S. cerevisiae*, splicing factors were detected, which underline the “intronrichness” of *S. pombe*.

RNA-Seq data of nascent RNA in comparison to nuclear and cytoplasmic mRNA show high correlation between samples. This underpins the high contribution of gene transcription to steady state mRNA levels. Average coverage profiles over intronless genes follow a 5′ to 3′ coverage gradient, which reflects the pool of nascent RNAs with varying lengths engaged in transcription by Pol II. This validates the purity of the nascent RNA preparation further.

Finally, “splicing per intron” values were calculated for the majority of introns in the three fractions and high co-transcriptional splicing levels are observed. It will now be interesting to understand how gene architecture and gene function contribute to the intron splicing patterns and the different preferences for co- and post-transcriptional splicing. This is subject of the next chapter in this thesis ([Chapter 4](#)).

CO-TRANSCRIPTIONAL SPLICING IN *S. POMBE*

4.1 GENE FUNCTION AND EXPRESSION

Half of all protein-coding genes from *S. pombe* carry at least one intron, but it remains unclear to which degree gene function and expression are associated with pre-mRNA splicing. In *S. cerevisiae*, for example, most of the intron-containing genes are very highly expressed and encode ribosomal proteins [Parenteau et al. 2011]. In *S. pombe* such functional distinction is not seen (own analysis), but might become apparent for groups with different splicing patterns. Therefore, I performed Gene Ontology enrichment analysis for the four groups with differential patterns of co- and post-transcriptional splicing identified in Section 3.2.3. Group I & II contain introns with very high intron splicing (group I highest co-transcriptional splicing) and intron splicing in group III is low and lowest in group IV.

The three most enriched GO terms for “Biological process” are shown in Figure 12. The term “mRNA cis splicing” is not only associated with very highly spliced genes, but also with genes, which contain lowly spliced introns indicating that many genes encoding for spliceosomal proteins are intron-containing. *S. pombe* cells were harvested in exponential growth, thus it is not surprising that terms associated with active metabolism and growth, for example “cytoplasmic translation”, “ribosome biogenesis”, “vesicle-mediated transport” and “protein targeting to ER”, are enriched in the two groups (I & II) containing highly spliced introns. Group III containing lowly spliced introns shows enrichment for terms associated with meiosis and mobilization of fatty acids, which are both not required in exponential growth and thus enrichment of spliced transcripts might not be necessary.

Steady-state mRNA expression levels are shaped by RNA transcription, degradation and processing. The high correlation between nascent RNA expression values and mRNA expression values in the *S. pombe* RNA-Seq data (Figure 10B) underlines the high contribution of RNA synthesis to steady-state mRNA expression values. In the following, I assess to which extent co- and post-transcriptional splicing levels are associated with high and low mRNA expression (Figure 13).

In the previous data analysis, pre-mRNA splicing was calculated for each intron (SPI). In order to assess a general correlation between pre-mRNA splicing and expression, gene splicing values were derived by averaging SPIs for individual genes. Scatter plot analysis of average gene splicing values and mRNA expression shows a very modest correlation ($R=0.14$ for nascent RNA gene splicing and $R=0.21$ for nuclear mRNA), which is not apparent for cytoplasmic mRNA ($R=-0.015$). Pearson correlation only allows to judge linear correlations, but non-linear forms of correlations of splicing and gene expression cannot be assessed.

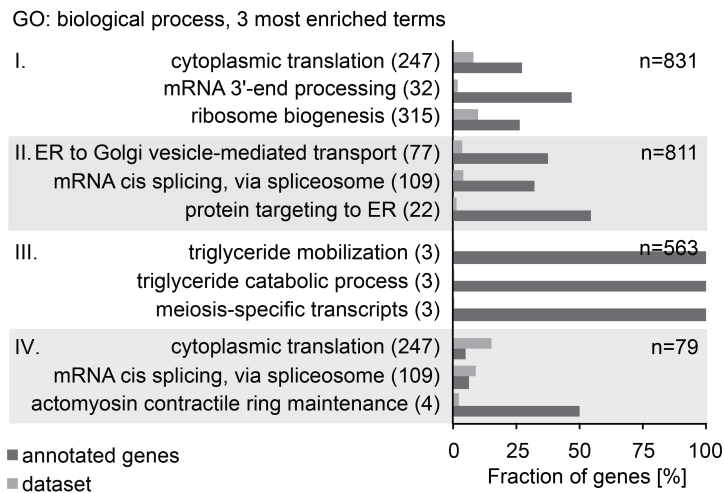


FIGURE 12 GO analysis of the four groups with differential splicing patterns. GO analysis was performed with topGO package in R. The three most enriched terms for “Biological process” are shown. Number of genes annotated for each term are given in brackets and numbers of genes per group are given in each barplot.

Grouping genes according their expression gives a second way of analyzing this large dataset and shows that the 40% highest expressed genes are significantly higher spliced co- and post-transcriptionally than lower expressed genes (Figure 13A). This trend is not seen for cytoplasmic steady-state mRNA splicing levels and suggests an interconnection between high transcription levels and pre-mRNA splicing. The reverse analysis of grouping introns according to their splicing values shows significant higher expression for genes with very high co-transcriptional splicing (group I, Figure 13B).

This could be a gene-specific characteristic and/or mRNA expression and pre-mRNA splicing are mechanistically linked. Efficient splicing might contribute to high gene expression and low splicing to lower gene expression by producing non-functional RNAs, which are degraded subsequently. Modulation of gene expression levels should provide more insight into the basis of the correlation between co-transcriptional splicing and intron-containing gene expression.

Therefore, *S. pombe* cells were treated with caffeine, a drug influencing the gene expression of hundreds of genes in a positive and negative manner [Rallis et al. 2013]. The comparison of mRNA expression levels and nascent RNA splicing levels upon caffeine treatment could then give an estimate to which degree pre-mRNA splicing shapes gene expression.

Upon caffeine treatment cells are known to show reduced proliferation and a gene expression program similar to a nitrogen starvation signature. Especially highly expressed genes, which are also highly spliced, involved in protein translation are known to be downregulated. Fifteen minutes of 10 mM caffeine treatment was considered optimal to detect immediate changes in gene expression and splicing (Section B.2, Figure 40).

Three replicates of nascent RNA and cytoplasmic mRNA were prepared (Figure 14A) and submitted for single-read Illumina sequencing (Section B.1 for mapping details and sample correlation). Raw reads were mapped to the *S.*

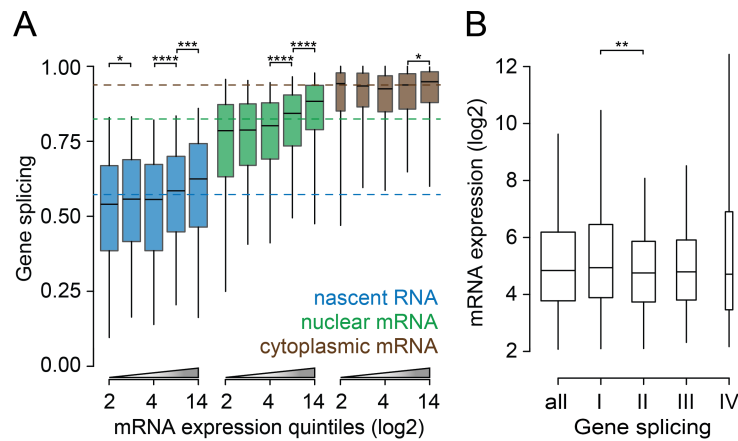


FIGURE 13 Co- and post-transcriptional gene splicing values correlate with mRNA expression. A: Boxplot of gene splicing values evenly grouped according to cytoplasmic mRNA expression (log2). B: Boxplot showing groups of introns with differential splicing patterns (Figure 11) and the associated mRNA expression distribution.

Asterisks indicate significance of direct neighbors according the Wilcoxon-rank sum test ($p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****); Boxplot whiskers correspond to 95% and 5% quantiles. Boxwidth is proportional to the square-roots of the number of genes per group.

pombe genome using Tophatz with the same settings as described in Section 3.2. Expression values and their significant differences were determined using Cufflinks and Cuffdiff. Splicing per intron (SPI) values were calculated as described in Chapter 3.

Overall, 1,512 genes were identified as being differentially expressed between water and caffeine treated samples (FDR-adjusted p -value < 0.05 , $FDR \leq 0.05$, Figure 14B). Out of those, 461 genes show expression changes by 2-fold or more. Downregulated genes show strong enrichment for GO terms (Biological process) associated with protein translation confirming published results that caffeine inhibits growth and proliferation. In addition GO term enrichment for amino acid-related processes is seen for upregulated genes confirming data that caffeine treatment mimics nitrogen starvation (Figure 14C). 566 of the differentially expressed genes are intron-containing and 553 of those had sufficient amount of data for splicing and expression correlation (Figure 14D). The difference in nascent RNA gene splicing and the log2-fold change in mRNA expression show a moderate Pearson correlation of 0.32 ($R=0.10$ for mRNA splicing changes/mRNA expression, $n=553$, Section B.2, Figure 41C). Almost no correlation ($R=0.12$) is observed for nascent RNA splicing values and nascent RNA expression, which allows to exclude that the observed positive correlation between nascent RNA splicing and mRNA expression originates from enhanced transcription, but might suggest that independent of potential changes in intron splicing, also degradation of unspliced transcripts could be enhanced with expression (Figure 41B).

Overall, high gene expression correlates with high co-transcriptional splicing levels. This could be a gene-specific property and/or a general principle in gene expression regulation. Changes in mRNA expression lead to modest co-trans-

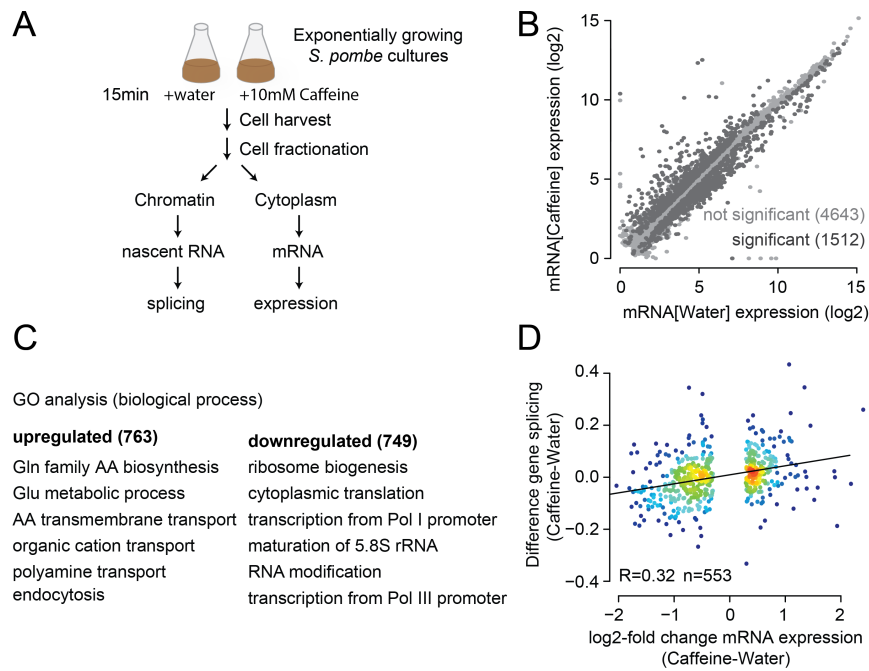


FIGURE 14 Gene expression changes correlate weakly with co-transcriptional splicing changes. A: Experimental outline to induce changes in gene expression upon caffeine treatment in *S. pombe* cells. B: Correlation of mRNA expression values in the two conditions identifies 1,512 differentially expressed genes (FDR-adjusted p-value < 0.05, FDR <= 0.05; 566 of those are intron-containing). C: GO term analysis of up- and down-regulated genes. The six most significantly enriched terms for “Biological process” are given; Pol - RNA Polymerase, AA - amino acid. D: Weak positive Pearson correlation between changes in nascent gene splicing and mRNA expression.

criptional splicing changes and no changes in final cytoplasmic mRNA splicing levels.

4.2 GENE ARCHITECTURE FEATURES OF DIFFERENTIALLY SPLICED INTRONS

S. pombe gene architecture shares common features with genes from other fungi species (e. g. intron length), but also with mammalian genes (e.g. exon length). The splice site recognition is assumed to happen through intron definition, which is similar to other fungi and organisms with short introns, e. g. *S. cerevisiae* and *D. melanogaster*, but different to organisms with primarily long introns, e. g. mammals. In the following, *S. pombe* co- and post-transcriptional splicing analysis with respect to the *S. pombe* gene architecture will be shown to address how differences in constitutive splicing are achieved.

Table 3 summarizes the main results. Genes carrying only a single intron and first introns in multi-intron genes are significantly lower spliced co- and post-transcriptionally compared to the global average (Figure 15A). Internal introns seem to be better spliced than the average intron. This is also reflected in the composition of intron groups, which were derived from euclidean clustering according to their co- and post-transcriptional splicing pattern (Figure 11). The

GENE FEATURE	CORRELATION TO PRE-MRNA SPLICING
intron position	first and single introns less well spliced internal introns highest splicing
number of exons	pre-mRNA splicing increases up to 6 exons
ORF length	with length better co-transcriptional splicing
CDS length	no correlation
distance TSS (internal & last introns)	with distance decrease in splicing
distance polyA site (first & internal introns)	no correlation
intron length	long (> 160 nt) & very short (< 40 nt) less well spliced
first exon length	short (< 200 nt) less well spliced
internal exon length	highest splicing ~25-100 nt
terminal exon length	< 300 nt less well co-transcriptionally spliced
exon GC-content	higher for high (co-transcriptionally) spliced introns
intron GC-content	lower for high (co-transcriptionally) spliced introns
splice site strength	less frequent splice sites and less splicing (trend, n.s.)

TABLE 3 Pre-mRNA splicing and gene architecture. Selected gene architecture features were correlated with co- and post-transcriptional splicing patterns using scatter plot analysis and grouping according to gene architecture feature or degree of splicing. Main results are summarized in this table.

fraction of internal introns in each group decreases with a decrease in pre-mRNA splicing, whereas the fraction of single and first introns increases (Figure 15B). Not only intron position influences pre-mRNA splicing, but also the length of the surrounding exons (Figure 15C, Figure 45, Section B.4). Short internal upstream and downstream exons are primarily found surrounding introns with high pre-mRNA splicing. No significant length dependence is seen for first and terminal exons and splicing, except that very short first exons (< 200 nt) and terminal exons (< 300 nt) are slightly less well spliced.

Most of the *S. pombe* introns are short and tightly distributed around the median length of 56 nt and only the “extreme” intron lengths (< 40 nt & > 160 nt) show lower co- and post-transcriptional splicing. Table 3 includes further gene architecture features (e. g. gene length and distance to transcript start and end), which have been analyzed and underline the significant correlation between pre-mRNA splicing, internal exon length and intron position within the gene.

Especially in the context of alternative splicing in higher eukaryotes, conservation of splice sites is often indicative for the amount of pre-mRNA splicing and the use of alternative splice sites. I determined the frequency of hexanucleotides at the 5′ end of introns (5′SS) and the frequency of trinucleotides at the 3′ end of introns (3′SS) and ranked them according their abundance in all annotated *S. pombe* introns. The fraction of the different splice sites does not change significantly among the groups of differentially spliced introns, indicating that splice site strength is not a strong determinant of splicing levels in *S. pombe* (Figure 16A-B). Intron and exon definition is facilitated by a bias in GC-content. Generally, introns have a lower GC-content (mean 0.30) than exons (mean 0.37). My analysis of intronic and exonic GC-content of the 4 groups with different splicing patterns shows that the difference in GC-content between introns and exons is strongest in group I, which contains introns with very high co-transcriptional splicing (Figure 16C).

Overall, links between pre-mRNA splicing levels and gene architecture are apparent and the predominant gene architecture of *S. pombe* supports high levels of co- and post-transcriptional splicing. The best spliced intron would be around 56 nt long, AT-rich, located within the gene and surrounded by short, GC-rich internal exons with a length ranging from 25 to 100 nt.

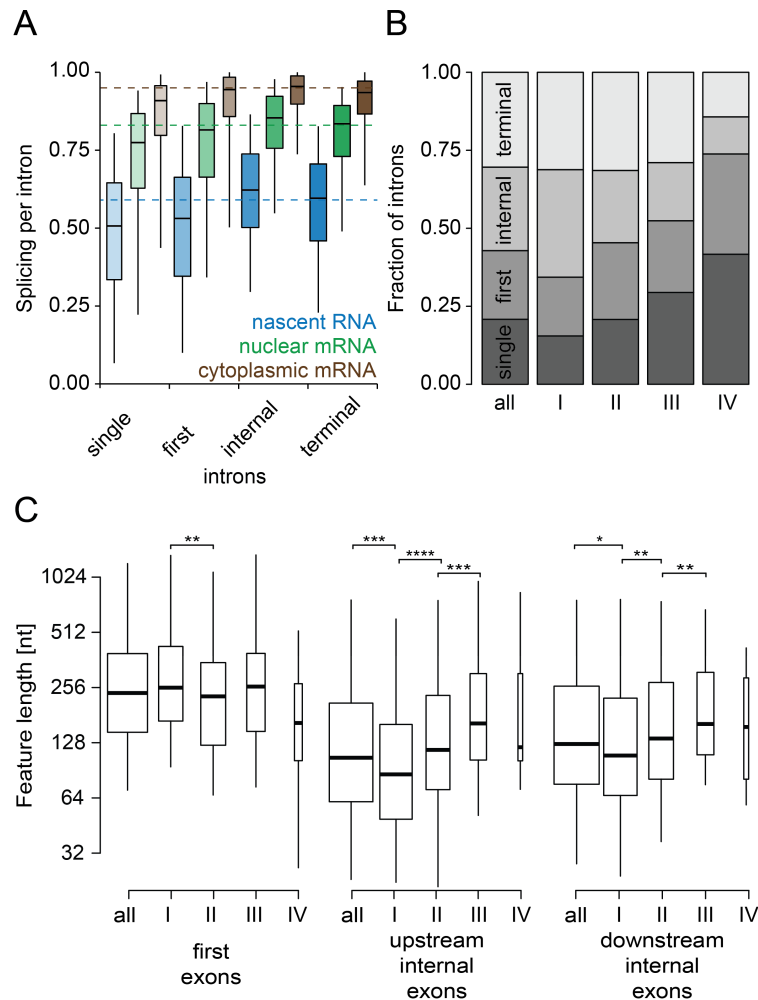


FIGURE 15 Differences in intron position and exon length associated with pre-mRNA splicing. A: First and single introns are spliced less well than internal and terminal introns. Boxplot of nascent RNA, nuclear and cytoplasmic splicing per intron values grouped according intron position (single-intron genes, multi-intron genes: first, internal and last). Median splicing of all introns shown as dashed line. B: Barplot showing groups of introns with differential splicing patterns (Figure 11) and the associated fraction of single, first, internal and terminal introns. C: Low intron splicing is associated with short first exons and long internal exons. Boxplot showing groups of introns with differential splicing patterns (Figure 11) and the associated first and upstream/ downstream internal exon length distribution.

Asterisks indicate significance of direct neighbors according the Wilcoxon-rank sum test ($p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****); Boxplot whiskers correspond to 95% and 5% quantiles. Boxwidth is proportional to the square-roots of the number of genes per group.

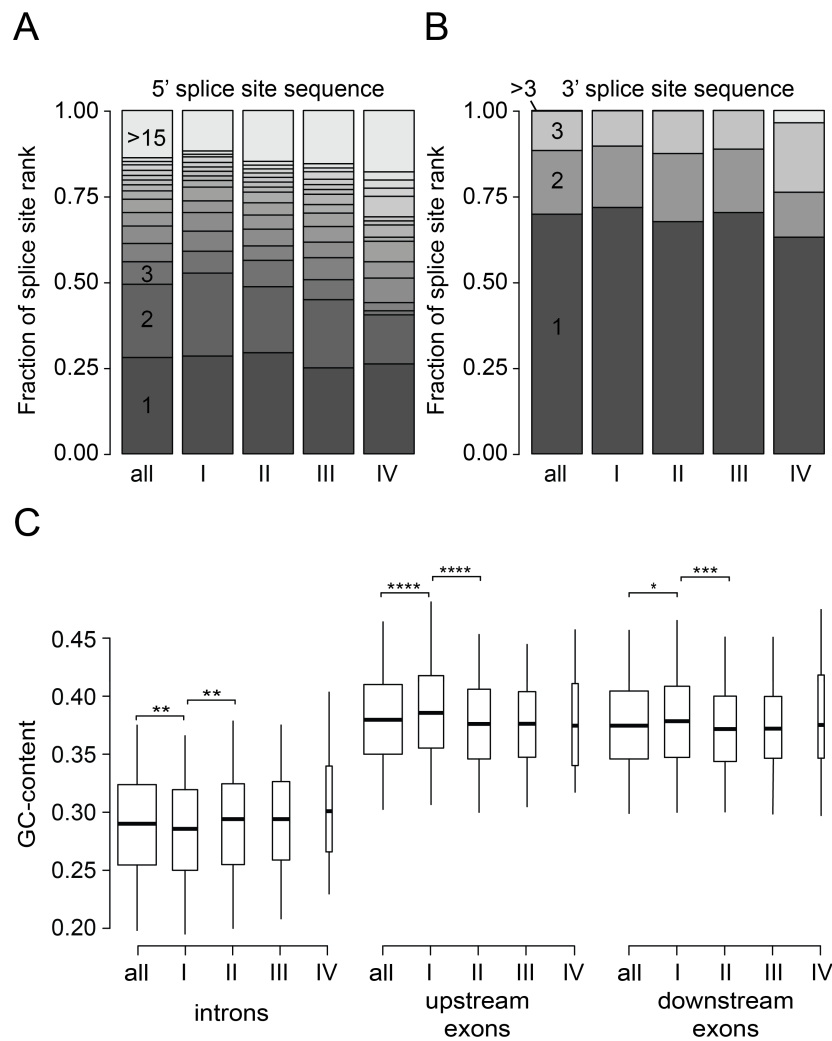


FIGURE 16 Differences in SS strength and GC-content associated with pre-mRNA splicing. A: Trend towards less abundant 5' SS sequence in lowly spliced introns. The first 6 nt of introns were ranked according to their abundance and the fraction of each hexanucleotide was determined for the groups of introns with differential splicing patterns (Figure 11). Fractions of 5' SSs with a rank greater 15 were pooled. B: Trend towards less abundant 3' SS sequence in lowly spliced introns. The last 3 nt of introns were ranked according to their abundance and the fraction of each trinucleotide was determined for the groups of introns with differential splicing patterns (Figure 11). Fractions of 3' SSs with a rank greater 3 were pooled. C: High intron splicing is associated with low intronic and high exonic GC-content. Boxplot showing groups of introns with differential splicing patterns (Figure 11) and the associated GC-content in introns, upstream and downstream exons.

Asterisks indicate significance of direct neighbors according the Wilcoxon-rank sum test ($p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****); Boxplot whiskers correspond to 95% and 5% quantiles. Boxwidth is proportional to the square-roots of the number of genes per group.

4.3 COUPLING OF INTRON SPLICING TO OTHER PROCESSING EVENTS

During and after transcription of protein-coding RNA multiple processing steps take place, which influence fate and function of the transcript. Shortly after synthesis of a few nucleotides of nascent RNA the 5' end gets chemically modified by addition of a 5' 7-methylguanosyl cap, which enhances stability of the transcript and serves as a landing platform for RNA-binding proteins throughout its presence in the nucleus and cytoplasm. During and after transcription pre-mRNA splicing takes place as highlighted by the presented data in the previous section. Other processing events, e.g. RNA editing, circularization of exons through backsplicing or transcript cleavage and polyadenylation towards the end of transcription could potentially influence each other and alter the gene expression outcome.

The following section focuses on the novel identification of circular RNAs in the *S. pombe* transcriptome generated through backsplicing and a yet unknown mechanism. Further, this section includes data on the order of intron removal in multi-intronic genes and establishes a link between splicing and transcript cleavage at the polyA site.

4.3.1 Appearance of circular RNAs

In order to assess whether formation of circular RNAs has an impact on co- and post-transcriptional splicing, I remapped my RNA-Seq dataset of chromatin-associated non-polyadenylated rRNA-depleted RNA using a transcriptome mapper optimized for detection of spliced and circular transcripts (segemehl 0.1.7 [Hoffmann et al. 2014]).

Circular transcripts were detected in addition to the identification of the majority of known splice junctions (75 circular RNAs present in all 3 replicates with unique junction Figure 17A). They could be grouped into two classes. Class I contains 45 circular RNAs, which circularization junction coincides with 5' and 3' ends of exons and are most likely products of backsplicing. They are similar in length to internal exons. The adjacent introns exhibit the known sequence logo of 5' and 3' SS (Figure 17A example I). This is not the case for the second class of identified circular RNAs, which are mainly derived from intronless ORFs and much shorter (Figure 17A example II). No junction sequence motif is apparent from the sequence logo analysis, which therefore points to a different splicing-independent process of generating those molecules.

Most of the identified circular RNAs are present to a much lower extent than the corresponding linear transcript. In order to assess the relative amount of circular reads to non-circular reads, I calculated the fraction of circular junction reads over all junction reads (junction read count from two independent spliced junctions divided by two to prevent overcounting of linear spliced junctions). The cumulative distribution of the fraction of circular reads shows that most circular RNAs are present at very low amounts (Figure 17A, right panel). Only 25% of the identified circular RNAs contribute to more than 5% of all junction reads at one locus.

Resistance to exonuclease digestion of circular RNAs can be used to validate their presence in cells [Wang et al. 2014]. Treated and untreated RNA samples are reverse transcribed with oligohexamer primers and cDNA is amplified by PCR or qPCR with inward primers picking up cDNA derived from linear and circular forms of RNA and outward facing primers exclusively targeting circular forms of RNA (Figure 17B, inset). I used this assay to validate two of the higher abundant circular RNAs derived from exons (SPBC345.06 and SPBC16G5.05c, Figure 17B, panel I) and two of the higher abundant circular RNAs from within ORFs (SPAC926.09c and SPAC1815.01, Figure 17B, panel II). In all 4 cases linear transcripts were efficiently degraded by RNase R, whereas circular transcripts remained stable under RNase R treatment in lysate and nucleus in three cases. The RNase R sensitivity for SPAC926.09c is lower than for linear transcripts, but very strong compared to the other three tested circular RNAs and thus indicates that a large fraction of this circular RNA is linearized within the cell after initial circle formation. Interestingly, all 4 tested circular RNAs were RNase R sensitive in the cytoplasmic fraction. Comparing the cytoplasmic or nuclear signal for circular RNAs encoded from ORFs suggests different compartment preference for the two species.

Pre-mRNA splicing, gene architecture, e.g. surrounding intron length, and sequence properties, e.g. GC-content, might facilitate circular RNA formation or enhance its stability. Indeed, circular RNAs of both classes show higher GC-content than exons (Figure 17C). Co- and post-transcriptional splicing are not significantly different between upstream and downstream introns or compared to all other introns (Figure 17D, panel I). However, downstream introns surrounding exons, which can form class I circular RNAs are usually longer than upstream introns (Figure 17D, panel II).

Taken together, I identified 75 circular RNAs using my nascent RNA-Seq dataset in *S. pombe*. I could group them into two classes according their way of formation (backsplicing, no backsplicing). They are not very highly expressed, have a similar minimal length of 43 and 37 nt, respectively, and out of 4 tested circular RNAs 3 were resistant to exonuclease activity. They exhibit a high GC-content and circular RNAs derived from backsplicing are similar in length to internal exons and are often adjacent to a longer downstream intron than upstream intron.

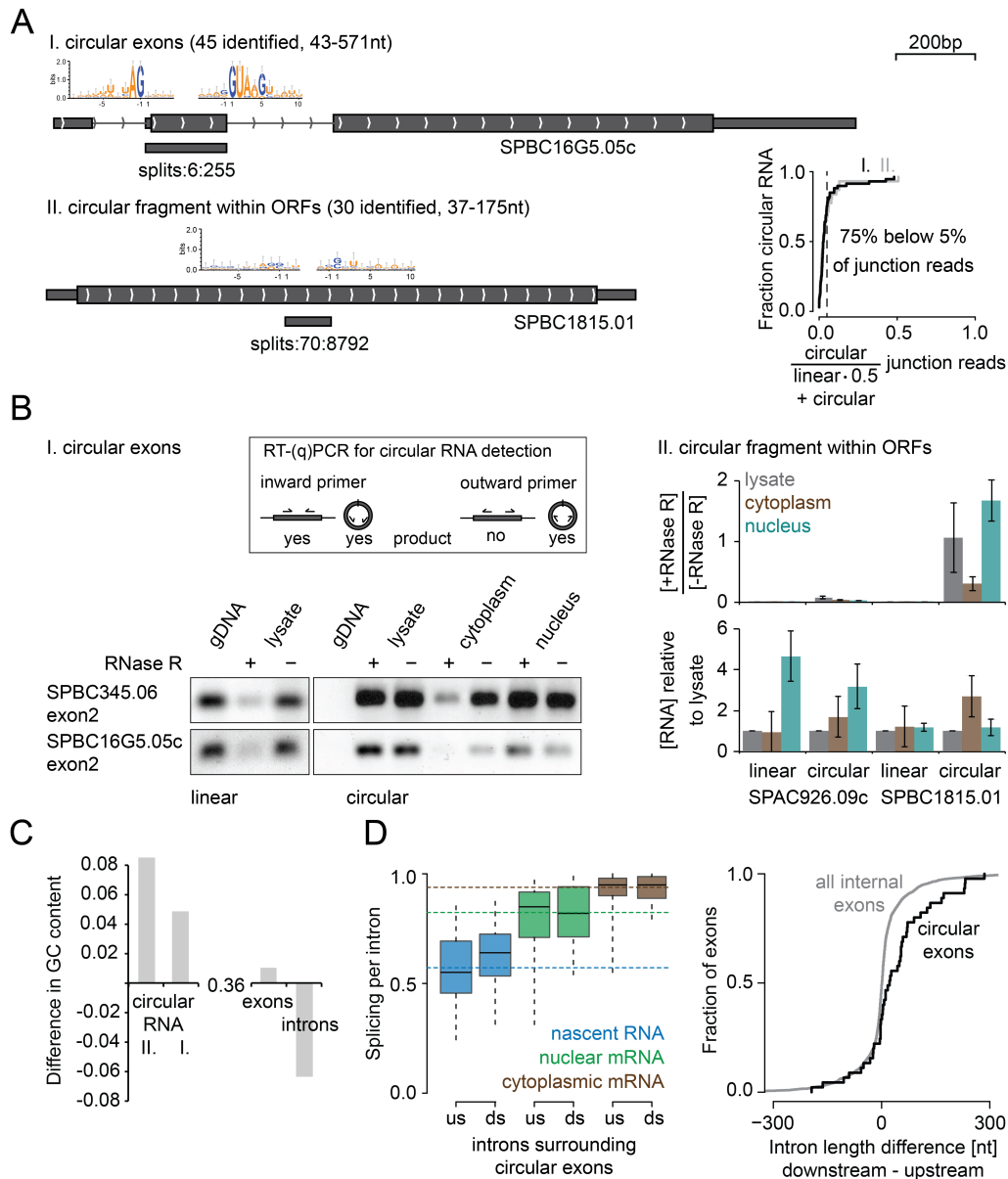


FIGURE 17 Circular RNA detection and characterization in *S. pombe*. **A:** Nascent RNA-Seq data were mapped with a transcriptome mapper optimized for detection of spliced and circular transcripts (segemehl 0.1.7 [Hoffmann et al. 2014]). 45 circular transcripts generated through backsplicing and bridging one or multiple exons have been identified in 3 replicates (most abundant example shown under I). 30 high-confidence circular fragments originating from within exons or intronless genes have also been detected. Most circular RNAs are low abundant compared to the corresponding linear transcript (cumulative curve of read fraction for the circular transcripts compared to the linear transcript). Split-read counts are given (splits:circular:linear). **B:** RNase R sensitivity for two examples per circular RNA class and their localization. Total, cytoplasmic or nuclear RNA were treated with RNase R (3'-5' exonuclease degrading linear RNA). RT-(q)PCR was done with random hexamers and PCR primers facing towards each other to amplify linear and circular species and primers facing away from each other to amplify only cDNA originating from circular RNA (n=2 (biological), SEM shown for RT-qPCR (2 biological with 3 technical replicates each)). 1 μ g of RNA was used for each experiment. **C:** Circular RNAs have very high GC-contents. Difference in GC-content relative to the genome average is shown. **D:** Circular exons are surrounded by normally spliced introns (us-upstream, ds-downstream; no significant difference in Kolmogorov-Smirnov and Wilcoxon-rank sum test, 1st panel), but have long downstream introns (2nd panel). The length difference between us and ds introns is plotted for circular exons and all internal exons and is significantly skewed towards long downstream introns ($p < 0.05$, Kolmogorov-Smirnov test).

4.3.2 Long read sequencing to link co-transcriptional RNA processing events

To this point I quantified global co- and post-transcriptional splicing levels in *S. pombe* from RNA-Seq data and correlated them with gene function, expression and gene architecture features. I could identify multiple aspects, which are distinct between introns and genes harboring different splice patterns. For example, I could show that first introns in multi-intronic genes are often less well spliced. How is this connected to the splicing of further downstream introns? Are they not removed as well or spliced first? I detected a substantial increase in levels of pre-mRNA splicing from nascent RNA to nuclear and cytoplasmic mRNA, which can be a result of post-transcriptional splicing or degradation of improperly spliced transcripts. What is the fate of unspliced transcripts, e. g. are they properly cleaved at the polyA site to induce transcription termination?

Those questions cannot be answered with conventional RNA-Seq, which generates millions of short reads originating from fragments of RNA. Ideally, sequencing of full-length transcripts can give an idea how RNA processing events like splicing of neighboring introns or splicing and transcript cleavage are connected. PacBio sequencing can be used for this purpose. Most *S. pombe* transcripts are shorter than 2 kb and thus ideally suited for the preparation of CCS (circular consensus sequence) 250bp to 2 kb PacBio libraries. In principle any double-stranded cDNA of appropriate length can be subjected to the PacBio library preparation, in which hairpin adaptors are ligated to both ends of the DNA. I developed a strategy and protocol to prepare those libraries (Figure 18A, Figure 46A, Section B.5).

The starting material is nascent RNA prepared from chromatin (Section 3.2). The goal, to sequence full-length nascent RNA, necessitates the introduction of a common sequence at the 3' end to allow full-length reverse transcription and further requires a common 5' end sequence to generate double-stranded cDNA. I account for this by ligating a DNA adaptor to the 3' end of nascent RNA. This procedure was adapted from [Churchman and Weissman 2011]. Details of the method and optimization can be found in Section 5.1.1. The adaptor serves as template for reverse transcription of nascent RNA. A reverse transcriptase with template switching activity was included in the protocol. This enzyme adds five non-templated nucleotides to the 3' end of the cDNA, which in most cases corresponds to the nascent RNA 5' end. Oligonucleotides annealing to this 5 nt overhang and the 3' DNA adaptor sequence can prime a final low-cycle PCR, which generates the double-stranded cDNA (Figure 18B). Lastly, primers and no insert DNA are removed in a DNA purification step and 500 ng to 1 µg are handed in for sequencing.

In the course of a PacBio sequencing run the DNA polymerase passes each DNA molecule multiple times moving along the circular template and thus sequences each DNA multiple times. The circular consensus sequence (CCS) from those sequencing rounds forms a mostly accurate read-out of the underlying DNA sequence. CCS reads correspond to cDNA sequences carrying all adaptors. Adaptors and barcodes need to be removed in order to accurately map transcripts back to the genome. The various post-processing steps are depicted in Figure 18C

and Section B.5, Figure 46B. The fraction of processed reads and the mapping efficiency are given in Table 11. It is noteworthy that overall only 50% of the data produce high-quality mapping. Even though rRNA was depleted from the sample multiple short transcripts map to the spacer region at the rDNA locus indicating that rRNA precursors and cleavage end products are abundant in the nascent RNA sample and not depleted using the commercial kit. Those reads are mainly filtered out with my mapping quality cutoff and not included in downstream data analysis (Figure 46D & Table 11).

In order to assess the quality of my PacBio data, I compared them to my nascent RNA-Seq data. Similarities and differences of PacBio data and Illumina RNA-Seq data are shown in Figure 18C. First of all, length of sequenced nascent RNA spreads from 22 nt to 3630 nt with a median of 242 nt. Illumina sequencing reads originate from fragmented RNA and are limited by the set read-length in the system, e. g. 76 bp. Second of all, correlation between PacBio data and nascent RNA Illumina sequencing data from Section 3.2 is strong ($R=0.56$) indicating that the number of observed reads for genes sequenced on the PacBio platform reflects nascent RNA transcription levels. However, the number of reads is generally low (protein-coding gene: median=16, mean=8.9) and with this library preparation, targeting all non-polyadenylated, non-ribosomal RNAs associated with chromatin, only the most abundant (often short) transcripts are sequenced (Figure 46D). For example, the highest fraction of reads is derived from snoRNAs, which are highly abundant nuclear RNAs and shorter than 200 nt. Using RT-qPCR and cellular fractionation of *S. pombe* wild-type cells, I could not detect a significant enrichment of snoRNAs in the chromatin fraction compared to nucleus and nucleoplasm. This indicates that the strong signal in the PacBio data results from a preference of sequencing short DNA molecules over long DNA molecules rather than enrichment of snoRNAs on chromatin (Section B.5, Figure 47A). Parsing through the data reveals many regions in the genome with transcripts very similar to annotated snoRNAs suggesting that PacBio sequencing of non-polyadenylated RNAs could be used to detect and annotate short ncRNAs like snoRNAs (Figure 47B).

PacBio sequencing cannot only be used to detect yet unidentified transcripts, but my goal is it to study the connectivity between RNA processing events, e. g. splicing of introns in transcripts carrying multiple introns or transcript cleavage at the polyA site and intron splicing. The analysis of the data with regard to those two aspects will be subject of the next two subsections.

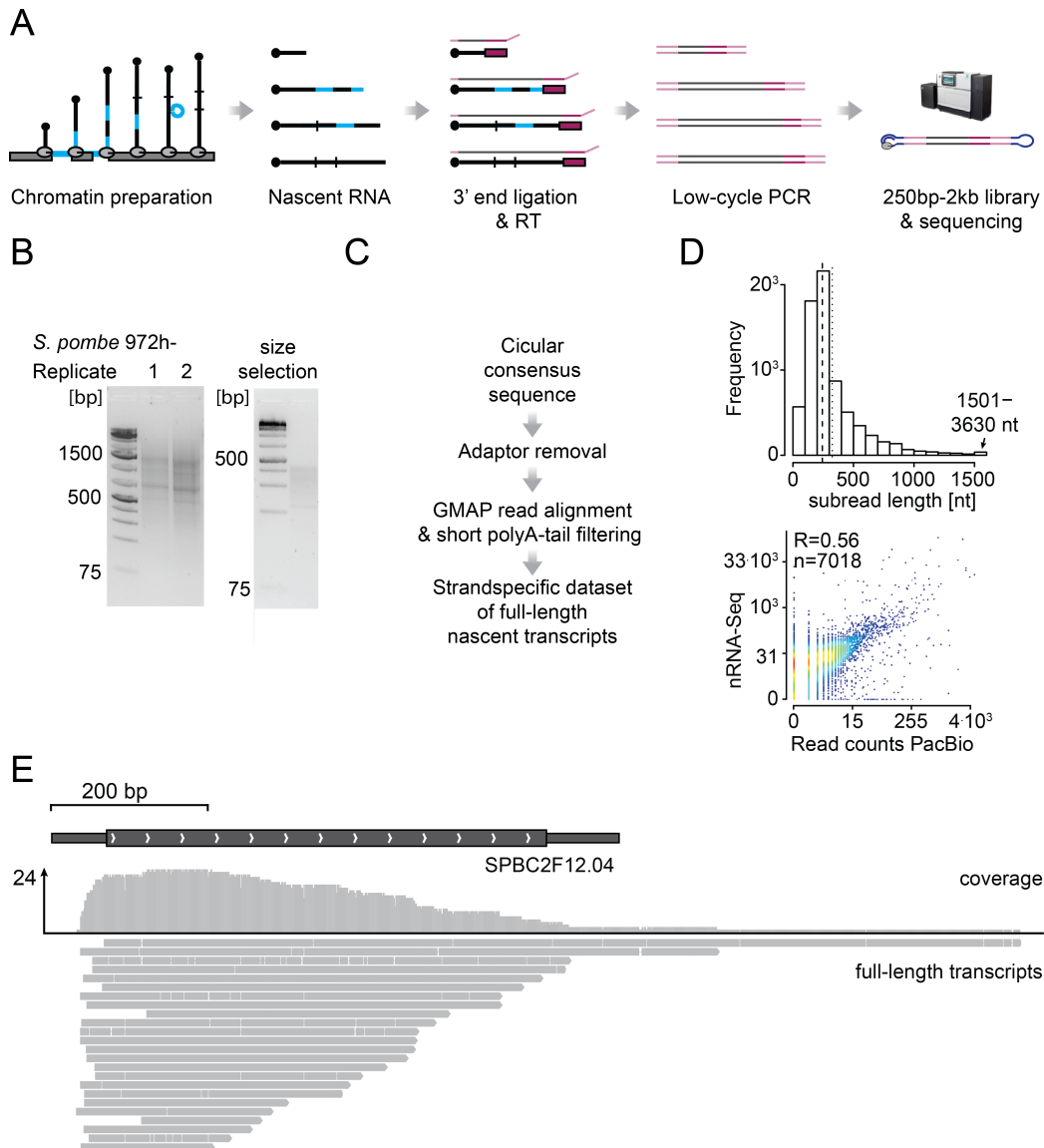


FIGURE 18 Long read sequencing library preparation and data processing. **A:** Nascent RNA from *S. pombe* is prepared from chromatin (2-intron gene with black nascent RNA attached via Pol II, the introns are highlighted in blue) and 3' end ligated to a DNA adaptor (purple box). The adaptor marks the 3' end of nascent RNA (the Pol II position during transcription) and serves as template for reverse transcription. Template switching reverse transcription adds a 5' adaptor, which serves as basis to generate double-stranded cDNA in a low-cycle PCR. Amplification-free PacBio library preparation and sequencing follows. All adaptors are colored in shades of purple and pink. **B:** Three *S. pombe* double-stranded cDNA libraries (2 full-length and 1 size-selected (< 500 bp)) are shown (1.5% Agarose gels, final double-stranded cDNA). **C:** Schematic of post-sequencing processing steps to remove adaptors, ensure strandedness and map transcripts back to the genome. **D:** The upper panel shows the cDNA length distribution without adaptors of the pooled data from **B** (Median=242 (dashed line), Mean=324 (dotted line), 16% of transcripts > 500 nt). The lower panel depicts the strong positive correlation between the gene expression values from short read Illumina nascent RNA-Seq data (Section 3.2) and the read-counts per gene of the full-length nascent RNA PacBio data. **E:** The 24 sequenced transcripts of one intronless gene are shown.

4.3.3 Order of intron removal

Half of all *S. pombe* intron-containing transcripts contain multiple introns (~1,400). The associated transcripts in my PacBio dataset covering multiple introns of one gene can be used to assess the order of intron removal. 2,499 transcripts (3.3% of all reads) from 688 genes fulfill this criterion. Four example genes are shown in [Figure 19A & B](#) and [Section B.6, Figure 49A & B](#). I grouped transcripts according to their splice pattern into three main groups:

1. all spliced: all introns have been removed already
2. all unspliced: all introns are still present in the transcript
3. partial: at least one intron is spliced and at least one other is not
 - a) in order: further separated according to how many introns are spliced upstream of an unspliced intron
 - b) not in order: further separated according to which intron is unspliced upstream of a spliced intron

[Figure 19C](#) shows how abundant the different transcript groups are and how many of them are detected per gene. Completely spliced transcripts (group 1) are most abundant (60% of all reads) and associated with 74% of the 688 genes harboring multi-intronic transcripts in my dataset. The second most prevalent transcript group is group 2 (completely unspliced) with 26% of all reads and in 53% of genes. Partially spliced transcripts are found in 15% of the reads and 25% of genes. Even though fewer partially spliced transcripts have been sequenced for fewer genes, the average number of transcripts per gene is similar with 2-3 reads/gene.

Partially spliced transcripts were further grouped according to the order of intron removal with 54% being spliced non-sequentially (not in order) and 44% sequentially (in order). Also 8 events of alternative splicing have been detected in the analysis (exon skipping or alternative splice site usage). Next, I asked for a prevalence in position of non-spliced introns in the “not in order” group and position preference of spliced introns in the “in order” group. This is depicted in the second pie-chart of [Figure 19C](#). First introns are most often not removed (76.4%) in “not in order” transcripts. This is in line with the lower co-transcriptional splicing rate measured by nascent RNA-Seq ([Figure 15](#)). Strengthening this point, for 155 transcripts containing the first intron, only 12 also contained further downstream introns (9x intron 2, 3x others). For the “in order” group of transcripts, most often the first intron only is spliced (87.8%) and includes also transcripts like the ones in example four ([Section B.6, Figure 49B](#)). There intron 2 is lower spliced and longer transcripts fall into the “not in order” group, which could very well be the outcome of the short transcripts classified as “in order”. The difference in co- and post-transcriptional splicing values of the 2nd and 3rd intron calculated from RNA-Seq data should give an idea how prevalent the outcome of not in order and in order splicing for this group of genes is. The difference is normally distributed with a mean of 0.02 for co-transcriptional splicing difference, suggesting that it is equally likely that the transcript introns will be spliced

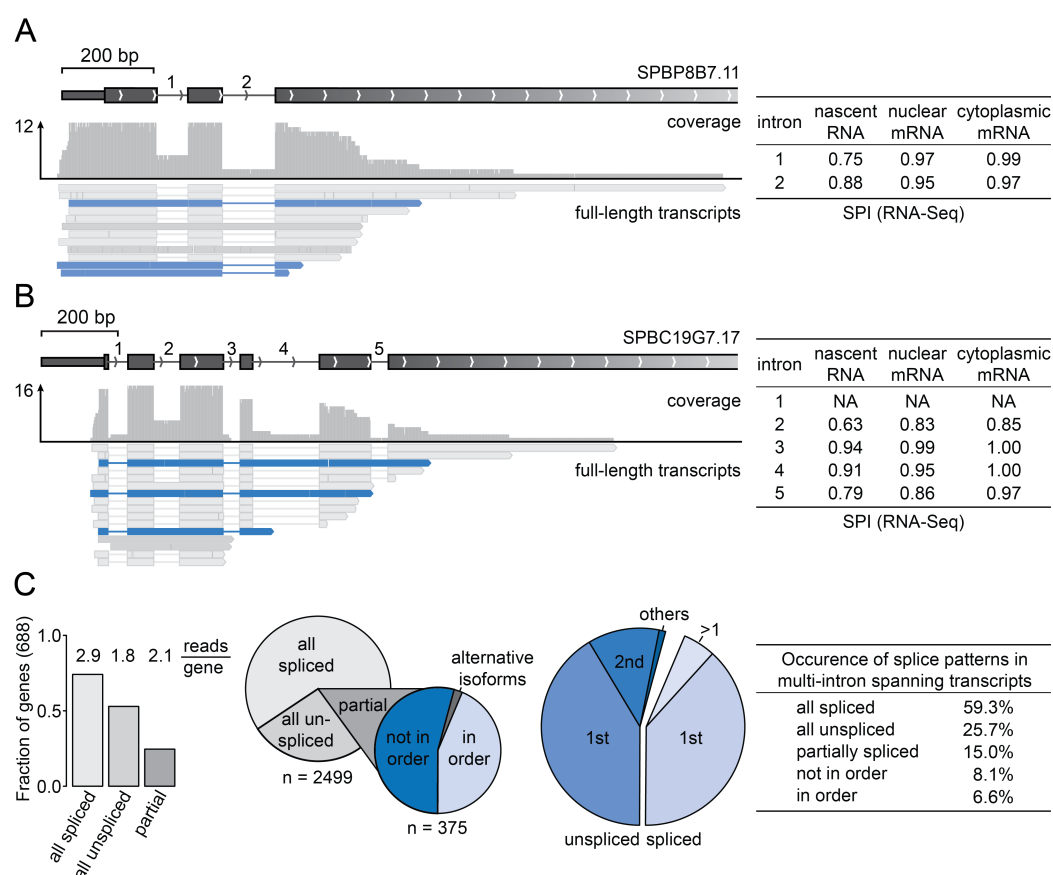


FIGURE 19 Connectivity of splicing of multiple introns. **A:** A two-intron-containing example gene with 12 sequenced transcripts is shown. All transcripts cover both introns. Most transcripts are completely spliced (light grey), two are completely unspliced (grey) and three transcripts show non-sequential splicing (blue) with no first intron splicing, but splicing of the 2nd intron. The table shows mean splicing values per intron (SPI) for nascent RNA, nuclear mRNA and cytoplasmic mRNA (Section 3.2). **B:** A five-intron-containing example gene with 16 sequenced transcripts is shown. All transcripts cover at least two introns. Most transcripts are completely spliced (light grey), two are completely unspliced (grey) and three transcripts show non-sequential splicing (blue) with the first and third intron being spliced, but no splicing of e. g. the 2nd intron. The table shows mean splicing values per intron (SPI) for nascent RNA, nuclear mRNA and cytoplasmic mRNA (Section 3.2). **C:** Barplot, pie charts and table summarizing the analysis of all multi-intron spanning transcripts of the PacBio dataset ($n=2,499$ from 688 genes). 74% of the genes have at least one fully spliced transcript (on average 2.9, light grey), 53% have at least one fully unspliced transcript (on average 1.8, grey) and 25% have partially spliced transcripts (on average 2.1, dark grey). The largest fraction of nascent transcripts is completely spliced (60%). 25% of transcripts are completely unspliced. The residual 15% of transcripts are partially spliced (partial) with 54% non-sequentially spliced transcripts (not in order, dark blue) and 44% sequentially spliced transcripts (in order, light blue). 8 alternative transcripts have been detected with skipped exons or alternative splice site usage.

in order or not in order for the group of 56 genes with more than two introns and splicing of the first intron (Section B.6, Figure 49C). In addition to the position preference of the non-spliced introns, also a length association is apparent. Introns surrounded by very short (upstream) internal exons (< 25 nt) are spliced to a lower extent (Table 3, Figure 45, Section B.6, Figure 49B).

Overall, most of the nascent transcripts are completely spliced, making up 60% of the reads, which is similar to the median intron splicing value derived from nascent RNA-Seq (Figure 11). Only few transcripts are partially spliced. Which introns are not spliced seems to be intron-specific and can be linked to intron position in the transcript or associated exon length in most of the cases.

4.3.4 *PolyA site cleavage and co-transcriptional splicing*

About one quarter of transcripts does not seem to be spliced co-transcriptionally (Section 4.3.3). What is the fate of those transcripts? Are they spliced post-transcriptionally to form mature mRNA or are they degraded? Many single gene examples can be found in the PacBio sequencing data, which indicate that unspliced transcripts are not properly cleaved at the polyA site. One example is shown in Figure 20A. Most transcripts are co-transcriptionally spliced and their 3' ends are before or at the polyA cleavage site indicative for nascent transcription and co-transcriptional transcript cleavage, however no splicing is observed for six transcripts, which are also not cleaved at the polyA site. The longest un-terminated non-spliced transcript is ~5x longer than the annotated ORF (Figure 20A). To see whether this holds true for other genes and transcripts in my dataset, I grouped transcripts according their 3' end position relative to the polyA site (+/- 20 nt or > 20 nt downstream) and calculated the fraction of spliced transcripts per group (Figure 20B, upper panel). Transcripts with their 3' end close to the polyA site are mostly spliced and longer transcripts are mostly unspliced for the > 500 transcripts originating from 184 genes. The 3' end distance to the polyA site varies for the detected transcripts with a peak 140 nt downstream of the polyA site and maximum detected distance to the polyA site of 2,157 nt (Figure 20B, lower panel).

This difference in transcript length and RNA processing depending on the splicing status of nascent RNA should be also visible in meta analysis. I aligned nascent 3' ends for intronless transcripts, intron-containing unspliced and spliced transcripts relative to the annotated polyA site and filtered for abundant snoRNA 3' ends, which would result in spikes masking the 3' end pattern of protein-coding nascent RNA. The resulting pattern is shown in Figure 20C. 3' ends from intronless transcripts have a broad distribution with a sharp drop at the polyA cleavage site. The 3' end count of spliced transcripts increases strongly towards the polyA site, peaks just before and immediately drops at the polyA site. The fewer unspliced transcript 3' ends do not show a strong peak over the gene body, but a modest decrease towards the polyA site and another small peak ~200 nt downstream of the polyA cleavage site representing the long non-terminated transcripts. Also long intronless transcripts are detected, but almost no spliced transcripts with 3' ends after the polyA site. The fraction of unspliced

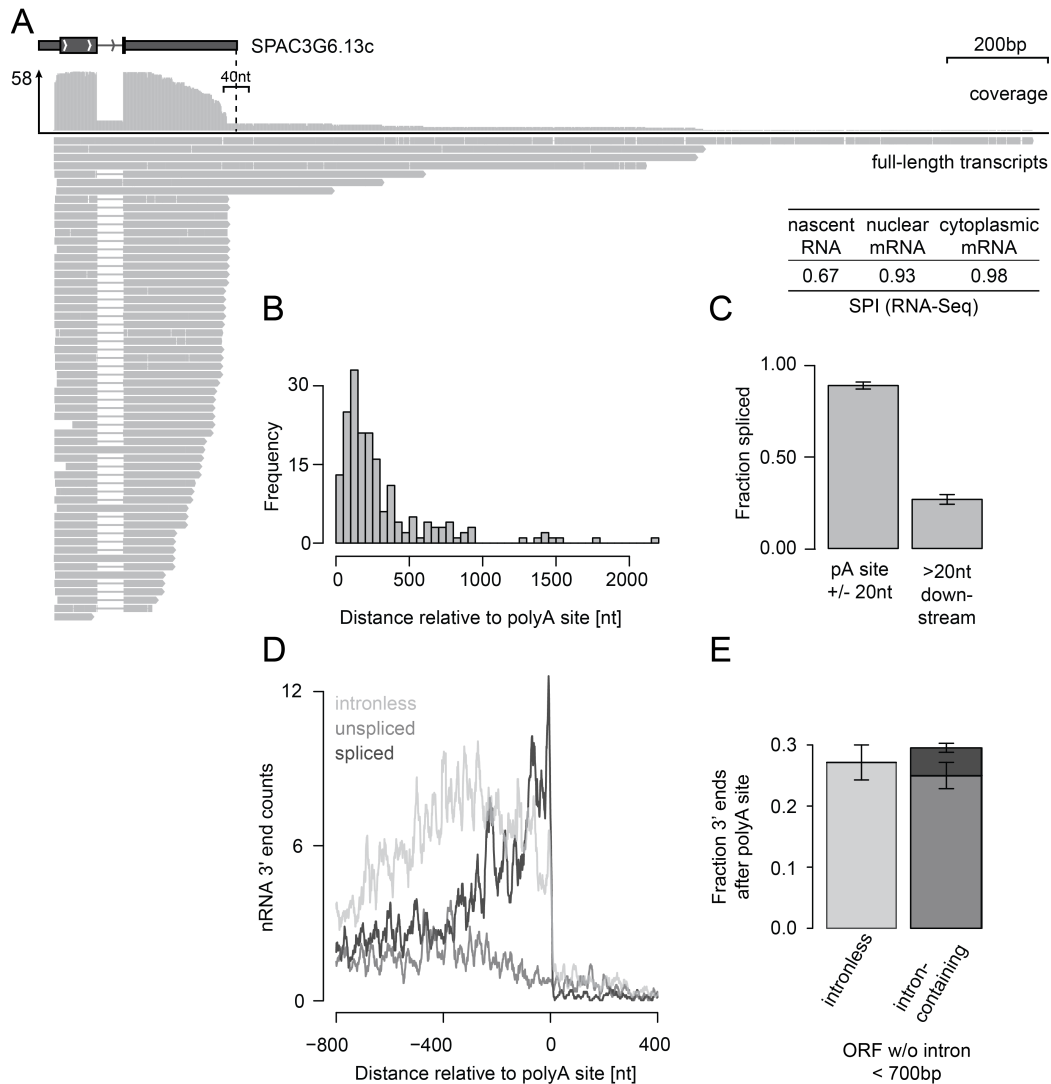


FIGURE 20 Co-transcriptional splicing and polyA site cleavage. **A:** A single-intron gene diagram with unspliced, non-terminated transcripts is shown. The intron is efficiently removed in most transcripts. Six unspliced transcripts are much longer than the annotated gene and not cleaved at the polyA site. The polyA site is highlighted with a dotted line and surrounded by the 40 nt window used in the analysis in **C**. **B:** Unspliced, non-terminated transcripts extend several hundred nucleotides over the polyA site. Histogram of 3' distances of those unspliced non-terminated transcripts relative to the polyA site with a peak at 140 nt, median of 202 nt and the maximum at 2,157 nt. **C:** Transcripts extending over the annotated polyA site (> 20 nt) tend to be unspliced opposite to transcript with 3' ends close (within a 40 bp window) to the annotated polyA site. Mean fraction spliced and standard deviation generated by bootstrapping (100 times) are shown in the right panel. **D:** Meta analysis of all 3' ends within -800 and +400 nt of the annotated polyA site for intronless, introncontaining spliced and unspliced transcripts. Summed read counts over position are shown, curves have been smoothed using a running average with a moving window width of 15. **E:** A small fraction of spliced non-terminated transcripts suggests a that intron splicing serves gene expression fidelity. The barplot shows the fraction of intronless transcripts extending over the polyA site in comparison to the corresponding fraction of unspliced and spliced introncontaining transcripts (stacked bar). Only genes with an annotated ORF smaller than 700 bp were included in the analysis (for introncontaining genes: ORF-introns < 700 bp). Mean and standard deviation are derived from bootstrapping (100 times). Colors correspond colors in **D**.

non-cleaved transcripts in single intron genes is similar to the fraction of intronless uncleaved transcripts in genes of similar size. However, the fraction of spliced reads is strongly depleted (Figure 20E). This suggests, a positive role for introns and its removal for transcription fidelity.

4.4 MULTIPLE LINKS BETWEEN PRE-MRNA SPLICING, GENE ARCHITECTURE AND EXPRESSION

Sequencing nascent and mRNA with RNA-Seq and PacBio sequencing allowed me to quantify global co- and post-transcriptional splicing levels (Section 3.2.3). The subject of this chapter was the analysis of intron and gene splicing values according to gene expression and gene architecture features and to characterize potential links between co-transcriptional splicing, expression and other co-transcriptional RNA processing events (Chapter 4).

Main results of this analysis are:

1. Most introns are efficiently spliced in *S. pombe* with a high fraction of co-transcriptional splicing (50% of introns are spliced to 58% and more co-transcriptionally).
2. Gene expression and co-transcriptional splicing are intrinsically correlated and there is a modest correlation of mRNA expression changes and changes in co-transcriptional splicing.
3. First introns are on average less well spliced and are the most common transcript class of not in order spliced multi-intronic RNAs.
4. Specific gene properties, e.g. short internal exon length and high difference in GC-content between introns and exons, can be linked to co-transcriptional splicing efficiency.
5. 60% of nascent transcripts are immediately spliced upon intron synthesis, 25% are completely unspliced and often not cleaved at the polyA site.
6. Circular RNAs are present in *S. pombe*, low expressed and not linked to observed pre-mRNA splicing patterns. Two classes of circular RNAs have been identified in intron-containing genes and intronless genes.

One aspect of the PacBio sequencing data has not been considered in this chapter. The 3' end of nascent RNA reflects the position of Pol II during transcription and thus the detection of spliced transcripts associated with the Pol II position can give an estimate on the position of co-transcriptional splicing during transcription. The data suggest that co-transcriptional splicing happens soon after intron synthesis is completed (Figure 20). However, the analysis is not quantitative and often convoluted in transcripts harboring multiple introns. In order to quantitatively investigate the progression of co-transcriptional splicing, I developed a paired-end deep sequencing protocol, called Single Molecule Intron Tracking (SMIT) for single intron genes in *S. cerevisiae*, where more details on

gene architecture and pre-mRNA splicing are known already and can be incorporated for data analysis. SMIT development and results will be subject of the next chapter ([Chapter 5](#)).

KINETICS OF CO-TRANSCRIPTIONAL SPLICING

5.1 A METHOD FOR SINGLE MOLECULE INTRON TRACKING ALONG WITH TRANSCRIPTION

Nascent RNA is attached to chromatin by Pol II and can be spliced during transcription. Illumina sequencing of nascent RNA 3' ends allows to determine the position of Pol II molecules during transcription along endogenous genes. Even though millions of nascent RNAs get sequenced, splicing and high-resolution transcription elongation information cannot be obtained on a single gene basis [Churchman and Weissman 2011]. 3' end density is too low in individual genes and the connectivity between splicing status and nascent 3' end is lost upon RNA fragmentation during library preparation. I developed a paired-end deep sequencing strategy (Single Molecule Intron Tracking, SMIT), in order to preserve Pol II position information and to connect this to intron splicing information. Targeting specific genes should also achieve high 3' end density over individual intron-containing genes, improving analysis. In SMIT defined ends of full-length cDNA from nascent RNA are sequenced. The 5' end read (SMIT read) informs about the splicing status and the 3' end read resembles the nascent 3' end. Being able to quantify both the splicing status and elongation status of single RNA molecules allows to determine, when splicing occurs relative to progress in transcription elongation.

The protocol itself consists of several steps, which are outlined in Figure 21. Nascent RNA is prepared from *S. cerevisiae* chromatin and depleted of polyadenylated RNA. The nascent 3' end is ligated to a DNA adaptor, SMIT adaptor, which serves as label for the Pol II position and also provides a common sequence to reverse transcribe nascent RNA into cDNA. Illumina sequencing generates many, but short sequence reads of the ends of longer cDNAs. Thus it is necessary to ensure that the sequenced region, e. g. 76 or 150bp, contains the information of interest. This can be achieved by placing a forward PCR primer in the first exon of *S. cerevisiae* genes in proximity to the intron to detect the splicing status of nascent RNA. This and a second PCR step allow the addition of further sequences and the barcodes that are required in an Illumina sequencing library. The nascent RNA preparation has been developed before for *S. cerevisiae* [Oesterreich et al. 2010] and forms the first block of the SMIT protocol. I optimized and adapted a commonly used 3' end ligation to the SMIT protocol requirements. In the following, data on the 3' end ligation optimization and SMIT PCR will be presented.

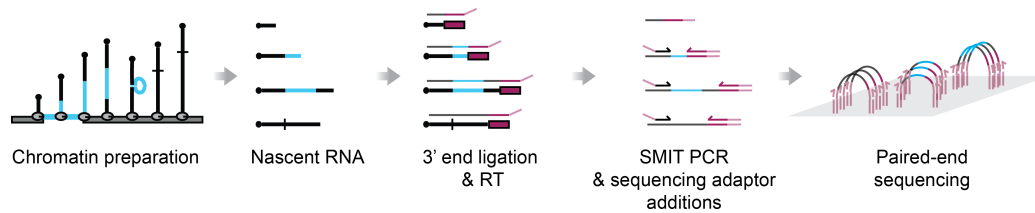


FIGURE 21 Single Molecule Intron Tracking (SMIT) library overview. Nascent RNA from *S. cerevisiae* is prepared from chromatin (single intron gene shown with Pol II (grey ball) and nascent RNA (black with blue intron) attached to it) and 3' end ligated to a DNA adaptor (SMIT adaptor, purple box). The adaptor marks the 3' end of nascent RNA (the Pol II position during transcription) and serves as template for reverse transcription. The SMIT PCR with a forward primer (gene-specific part black) placed close to the 5' SS defines the 5' end of the cDNA to be sequenced to identify the splicing status of each transcript. The reverse primer anneals to the 3' end adaptor sequence. A second PCR attaches final Illumina sequencing adaptors and barcodes (light pink). Illumina paired-end sequencing follows.

5.1.1 3' DNA adaptor ligation to preserve Pol II position during transcription

The 17nt DNA sequence used as SMIT-adaptor has been used before in profiling nascent RNA 3' ends [Churchman and Weissman 2011] and for miRNA sequencing library preparation [Lau et al. 2001]. I changed the design and included a five nucleotide random barcode to the 5' end to minimize sequence associated ligation biases and to be able to select unique molecules from the pool of amplified cDNAs after sequencing. The adaptor is pre-activated for ligation by 5' pre-adenylation and inactivated for ligation at the 3' end with a 3' dideoxynucleotide. A genetically improved T4 RNA ligase 2, truncated K227Q, is included in the protocol. SMIT adaptor modifications and the use of the optimized T4 RNA ligase ensure low side product formation [Viollet et al. 2011]. Optimization and quality tests of the 3' end ligation are shown in Figure 22.

I used short *in vitro* transcribed RNAs to assess the ligation efficiency and potential preference for certain RNA 3' ends. The RNA was ligated overnight at 16°C and subsequently analyzed by 10% or 15% denaturing TBE-Urea polyacrylamide gel electrophoresis (PAGE) (Figure 22A). The bottom bands resemble unligated RNA and the top band ligated RNA carrying the 3' end adaptor. Two different RNAs show very different ligation efficiencies, which also depend on the concentration of the adaptor. There is no RNA 3' nt preference using my adaptor design (Figure 22B). The differences in ligation between different *in vitro* transcribed RNAs (Figure 22A) can be significantly reduced by enhancing molecular crowding through PEG 8000 addition (Figure 22C). The final reaction conditions of 25% PEG 8000, a high adaptor to RNA ratio (20:1) and heat denaturation of RNA ensure efficient (70%) and uniform (standard deviation 4%) ligation of 7 different *in vitro* transcribed RNAs.

For nascent RNA a complete shift towards higher molecular weights is observed in denaturing TBE-Urea PAGE after ligation, which is indicative for efficient ligation of all nascent RNAs (Figure 22D). 3' end ligated nascent RNA can thus be

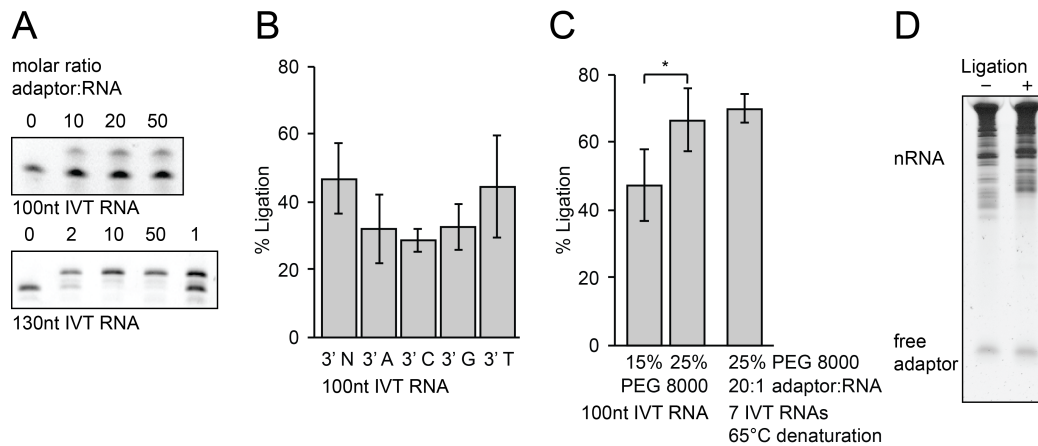


FIGURE 22 Characterization of the 3' end SMIT adaptor ligation. **A:** SMIT adaptor ligation increases with higher adaptor concentration and varies between different RNAs. Two *in vitro* transcribed (IVT) RNAs (100 nt and 130 nt part of human 18S rRNA) were ligated to varying amounts of 3' end DNA adaptor and analyzed by 10% Urea-PAGE. The lower band corresponds to the unligated RNA and the upper band resembles the ligated product. **B:** Quantification of 3' end ligation of the 100 nt IVT RNA with varying 3' ends. No significant difference in ligation is observed for the 4 possible end nucleotides and a random 3' end with two adaptor concentrations ($n=6$, SD is shown, pooled samples from 10:1 and 20:1 adaptor:RNA ratios). **C:** Enhanced molecular crowding facilitates 3' end ligation. Quantification of 3' end ligation of the 100 nt IVT RNA with 15% or 25% PEG 8000 ($n=5$ and 4, SD is shown, $p < 0.05$ two-sided Student's *t*-test) is presented in the first two bars. 3rd bar shows 70% average ligation efficiency for 7 different IVT RNAs with optimized reaction conditions (SD is given). **D:** 3' end ligation for nascent RNA is very efficient. All nascent RNA species detectable in the gel are shifted towards higher molecular weights (10% TBE-Urea PAGE).

used as a template for reverse transcription (RT) using the adaptor sequence for priming.

5.1.2 SMIT PCR and library to identify splicing position

The development of a novel deep sequencing approach, like the targeted SMIT approach, requires testing of each step within the protocol to ensure that the resulting libraries represent the genes of interest. First, I tested for specificity of the SMIT PCR. The forward primer placed at the end of the first exon of three *S. cerevisiae* genes is combined with a reverse primer binding to the transcript's last exon. Thus only one band is expected in the genomic DNA control, but depending on the fraction of co-transcriptional splicing one or two bands are expected for a PCR on cDNA from 3' end ligated nascent RNA. This is indeed the case (upper panel of [Figure 23A](#)) and almost no background amplification is detected in the various controls. Second, I replaced the gene-specific PCR reverse primer with the 3' end adaptor sequence to test, if the SMIT adaptor sequence can be used as specific reverse primer binding site. Indeed, I obtained a gene-specific product smear in the SMIT PCR, but not in the different controls ([Figure 23A](#), lower panel).

For additional 24 SMIT forward primers similar gene-specific cDNA smears can be observed (Figure 23B). Afterwards the gene-specific SMIT PCR samples are pooled, purified and submitted to a second PCR with a limited, experimentally tested number of cycles. Often a “no insert product” is observed, which runs at 150 bp and carries only adaptors (Figure 23C, first lane). PCR product purification using AMPure beads at a bead volume of 0.8 μ L per 1 μ L PCR solution reduces the amount of “no insert product”. In this procedure DNA molecules of defined sizes are precipitated with high salt and molecular crowding on paramagnetic beads. Gel-based size selection prior to sequencing for double-stranded cDNA greater 150 bp further enhances the amount of correct product.

I designed the SMIT libraries in a way that Illumina barcodes can be included and multiplexing of replicates and different samples is possible per sequencing lane. Furthermore, I included a second random barcode just downstream of the 3' end adaptor. This ensures that the first five bases sequenced are random, and it allows for proper DNA cluster identification during sequencing (Figure 50A). Paired-end sequencing with 76-100 nt reads has been done *S. cerevisiae* genes shown as SMIT PCR examples in Figure 23B (agarose gel images of 6 replicates in Figure 50B). In order to counteract the depletion of short nascent RNAs due to the required size-selection, I prepared a second, size-selected library for very short transcripts (25-250 nt) including the same genes (Figure 50C). This initial dataset was extended to 88 genes to obtain a broad spectrum of endogenous splicing kinetics in *S. cerevisiae* (Figure 50D-E).

In order to obtain the information of Pol II position and splicing from sequencing data, I developed a processing pipeline, which selects read pairs carrying the 3' end adaptor, removes PCR duplicates and then maps the high-confidence SMIT read (splice status) to the set of spliced and unspliced junctions and the 3' end read (Pol II position) back to the genome (Section C.1). The raw, processed and mapped read counts are shown in Table 12. The individual samples and replicates correlate well among each other with an average Pearson correlation of 0.64 for the long RNA samples and 0.75 for the short RNA samples (Section C.1, Figure 51A). 3' end counts per nucleotide position were correlated among replicates. In total 26.3 Mio reads were mapped to 217,145 unique chromosome positions. Most positions are only covered by one read. The median of the read count/position distribution is 5 and the average is 109 reads per position (Section C.1, Figure 51B).

Splicing analysis follows data processing and is shown in the next section (Section 5.2).

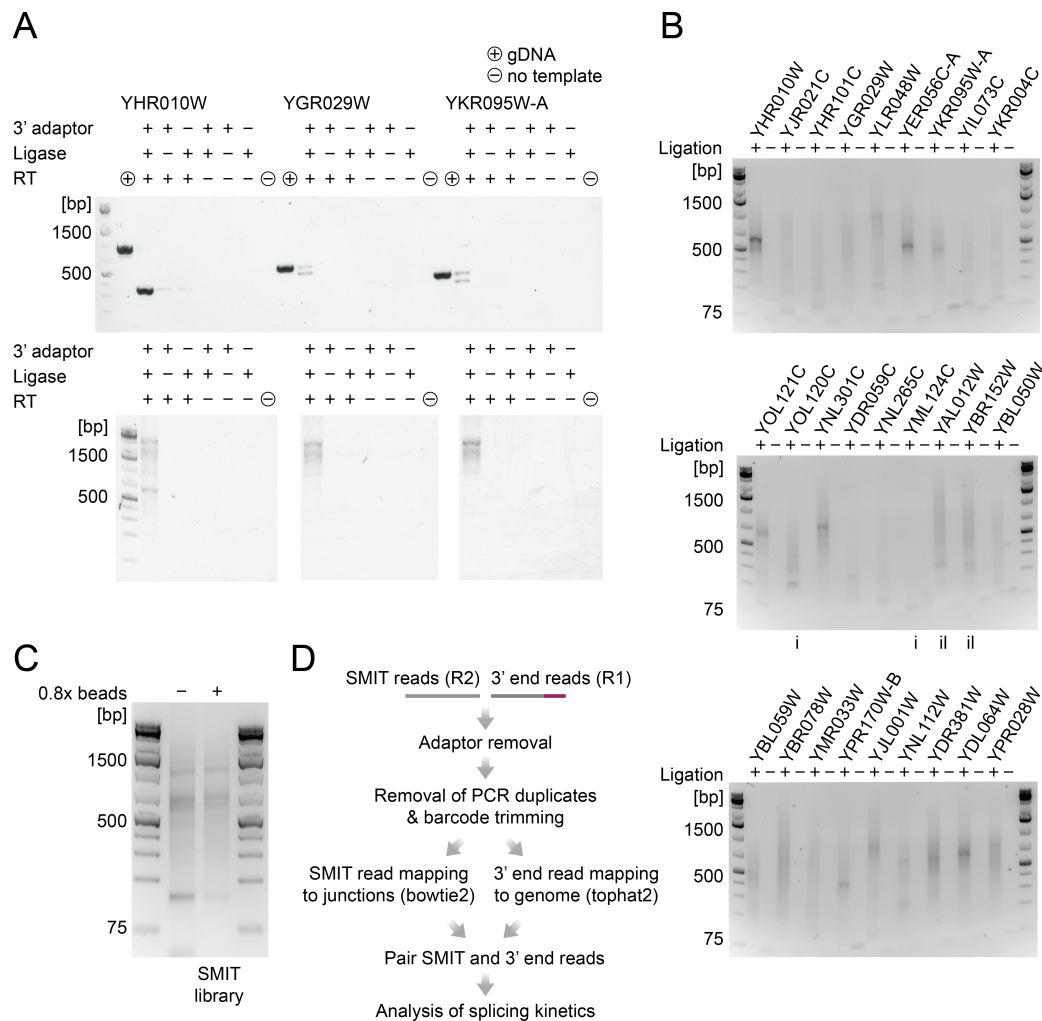


FIGURE 23 SMIT PCRs and library. **A:** Agarose gel electrophoresis showing the SMIT PCR results with gene-specific reverse primer (defined products, upper panel) and reverse 3' end adaptor primer (cDNA smear, lower panel). Minus ligation and minus reverse transcription controls show no or very low background amplification. **B:** SMIT PCR for 27 genes with forward primers in the first exon, intron (i) or intronless ORF (il) and reverse 3' end adaptor primer produce gene-specific cDNA smear. Similar to lane 1 & 2 of lower panel in **A**. **C:** cDNA smear originating from pooled SMIT PCRs and 2nd PCR, which attaches Illumina adaptors and barcodes. PCR artefacts and residual primer molecules are removed by 0.8x volume of AMPure beads and the final library consists of cDNA ranging from ~200 bp to ~2 kb. **D:** Overview of post-sequencing processing steps to identify high confidence 3' end reads and associated splice junction reads (SMIT read). Further details are given in [Figure 50](#).

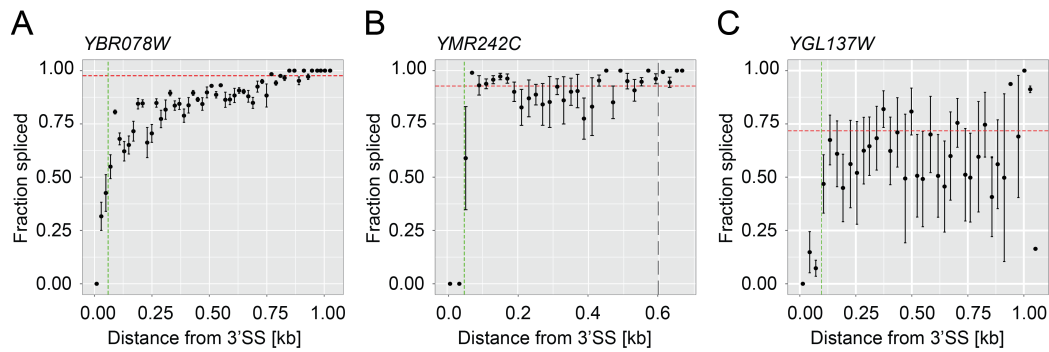


FIGURE 24 Onset and progression of co-transcriptional can be monitored by SMIT. Three SMIT traces showing different patterns of co-transcriptional splicing are given in A-C. Fraction of splicing is plotted with regard to distance to the 3' SS. The dashed red line indicates co-transcriptional splicing saturation. The green line marks the 50% saturation position and the dashed black line (in B) marks the terminal exon end. Data points represent binned means (30 nt) with SD from pooled data of all replicates. Data visualization in Figures A-C and the analysis of the processed and mapped SMIT data were done by Fernando Carrillo Oesterreich. More SMIT traces are given in [Figure 52](#).

5.2 THE POSITION OF CO-TRANSCRIPTIONAL SPLICING IN *s. cerevisiae* GENES

5.2.1 Common and unique characteristics of co-transcriptional splicing in individual genes

The next part in analysis involves pairing the splice status information with the associated 3' end information. In the SMIT datasets the number of observed 3' ends decreases exponentially with insert length. This provides a challenge in accurately quantifying the relative amounts of spliced and unspliced transcripts at one position, because spliced transcripts are shorter than the respective intron-containing unspliced transcript. Read count normalization according to their exponential distribution in insert length has been performed for this reason. Subsequently, the fraction of spliced transcripts per position can be computed ([Section 9.19](#)).

[Figure 24](#) shows three examples of SMIT traces (more examples in [Section C.2](#), [Figure 52](#)). The fraction of splicing per position is plotted relative to the end of the respective intron. To be robust against noise, genomic positions were grouped into 30 nt bins and the mean and standard deviation of the co-transcriptionally spliced fraction is shown. No splicing can be detected at the end of an intron (3' SS), but within a few nucleotides the fraction of spliced transcripts increases dramatically. This increase is almost stepwise for examples 2 and 3, but more gradual for example 1. All example traces show saturation, albeit at different levels. The saturation value represents a measure of co-transcriptional splicing fraction reached before termination. To compare co-transcriptional splicing patterns between genes, the saturation values (dashed red line) and the 10%, 50% (dashed green line) and 90% saturation position are used ([Section 9.19](#)).

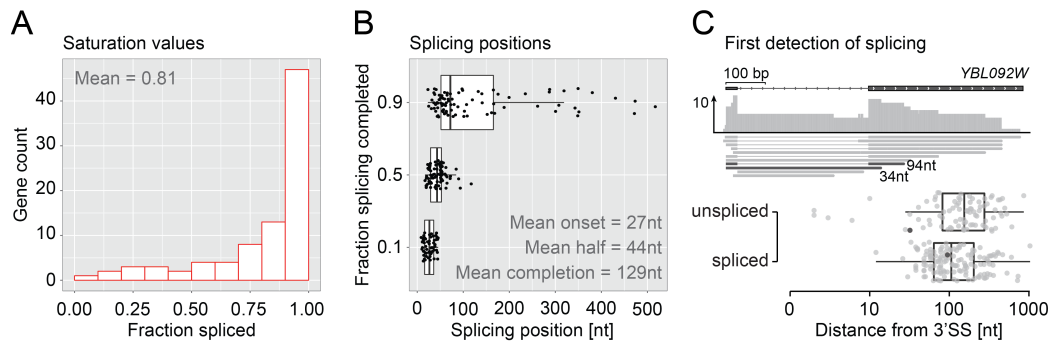


FIGURE 25 Co-transcriptional splicing is efficient and fast. A: Distribution of co-transcriptional splicing values at the point of transcription termination. The histogram plotting gene count versus the saturation fraction spliced shows that most genes are almost completely co-transcriptionally spliced. B: Position relative to the 3' SS where 10%, 50% and 90% of the saturation value is reached. C: Single gene example of *S. cerevisiae* PacBio data confirming detection of splicing close to the 3' SS. All intron-spanning reads were grouped into spliced (138) and unspliced (85) and the position distribution of the closest 3' end relative to the 3' SS is plotted. The black dots highlight the 3' ends of the first (un)spliced read of the example gene. Figures A-B and the analysis of processed and mapped SMIT data were done by Fernando Carrillo Oesterreich.

5.2.2 Frequent and rapid co-transcriptional splicing for most assayed genes

A histogram for saturation values and distributions of positions for 10%, 50% and 90% saturation are shown for 88 genes in Figure 25. Most assayed *S. cerevisiae* introns are spliced almost completely co-transcriptional with a mean of 0.81. Although the saturation varies from almost no co-transcriptional splicing to 100% co-transcriptional splicing, there seems little variation in how soon after intron synthesis the saturation in co-transcriptional splicing is reached (Figure 25B). The majority of co-transcriptional splicing happens within the distance of 100 nt after the intron end and thus in close vicinity to the transcribing polymerase and it starts around 27 nt downstream of the 3' SS (10% of saturation). *S. cerevisiae* PacBio data support the notion of co-transcriptional splicing close to the 3' SS. I prepared two *S. cerevisiae* PacBio datasets as described for *S. pombe* (Section 4.3.2, Section 9.11) and analyzed them with regard to the 3' end distance to 3' SSs of intron-spanning transcripts. Closest 3' ends of unspliced transcripts are normally distributed downstream of 3' SSs (minimum 2 nt downstream of 3' SS, Kolmogorov-Smirnov test $p > 0.05$ compared to simulated normal distribution). Closest 3' ends of spliced transcripts follow a different distribution ($p < 0.01$) and the first transcript is only detected 12 nt downstream of the 3' SS (Figure 25C). This suggests that there is a minimal distance between the 3' SS and the onset of splicing.

5.3 FAST CO-TRANSCRIPTIONAL SPLICING IS WIDESPREAD

So far I described the development of a paired-end deep sequencing method to measure co-transcriptional splicing with distance to the assayed intron. The re-

sulting data give an idea, where RNA polymerase resides while splicing occurs. Overall the data show that co-transcriptional splicing happens within 100 nt. Afterwards saturation is observed, which means that the fraction of spliced transcripts and unspliced transcripts for the assayed genes does not change anymore before transcription termination.

Compared to results from other studies, co-transcriptional splicing occurs much faster and closer to the end of an intron than previously estimated (Section 1.1). I performed a SMIT-like PacBio experiment on a splicing reporter transgene, for which spliced products have been only detected 500 nt downstream of the 3' SS. This produced similar results as obtained for endogenous genes by SMIT and suggests that co-transcriptional splicing is as sudden and close to 3' SS as for fast splicing endogenous genes (Figure 26). For the transgene carrying the *S. cerevisiae* consensus 5' SS (GUAUGU) 32% of transcripts shorter than 200 nt were spliced, in a 5' SS mutant (U₄C) only 15% showed co-transcriptional splicing in this window. This is similar to a later spliced endogenous gene carrying a non-consensus branchpoint sequence (Figure 26B-C, Section C.2, Figure 52 example 7).

One concern could be that degraded/hydrolyzed RNA is ligated to the 3' end DNA adaptor. In a 3' end ligation test comparing commercially synthesized 31 nt 3' hydroxylated and 3' phosphorylated RNA (hydrolyzed RNA would be 3' phosphorylated) I could not detect any ligation by denaturing Urea-PAGE analysis (Section C.3, Figure 54C). I also assessed 3' end ligation after alkaline hydrolysis of both example RNAs. The pool of hydrolyzed 3' hydroxylated is still ligatable, but the pool of 3' phosphorylated RNAs is not. This confirms that RNA 3' ends after hydrolysis are 3' phosphorylated and thus cannot be ligated (Figure 54A-C). Hence, the SMIT experiment should be specific to nascent RNAs and not detect RNA degradation intermediates.

In conclusion, this chapter on the position of *S. cerevisiae* co-transcriptional splicing and the results on gene architecture features associated with *S. pombe* co-transcriptional splicing from Chapter 4 give evidence that the majority of pre-mRNA splicing is co-transcriptional in the two yeasts and that co-transcriptional splicing happens within 100 nt downstream of the 3' SS with a small delay of ~27 nt after the 3' SS synthesis.

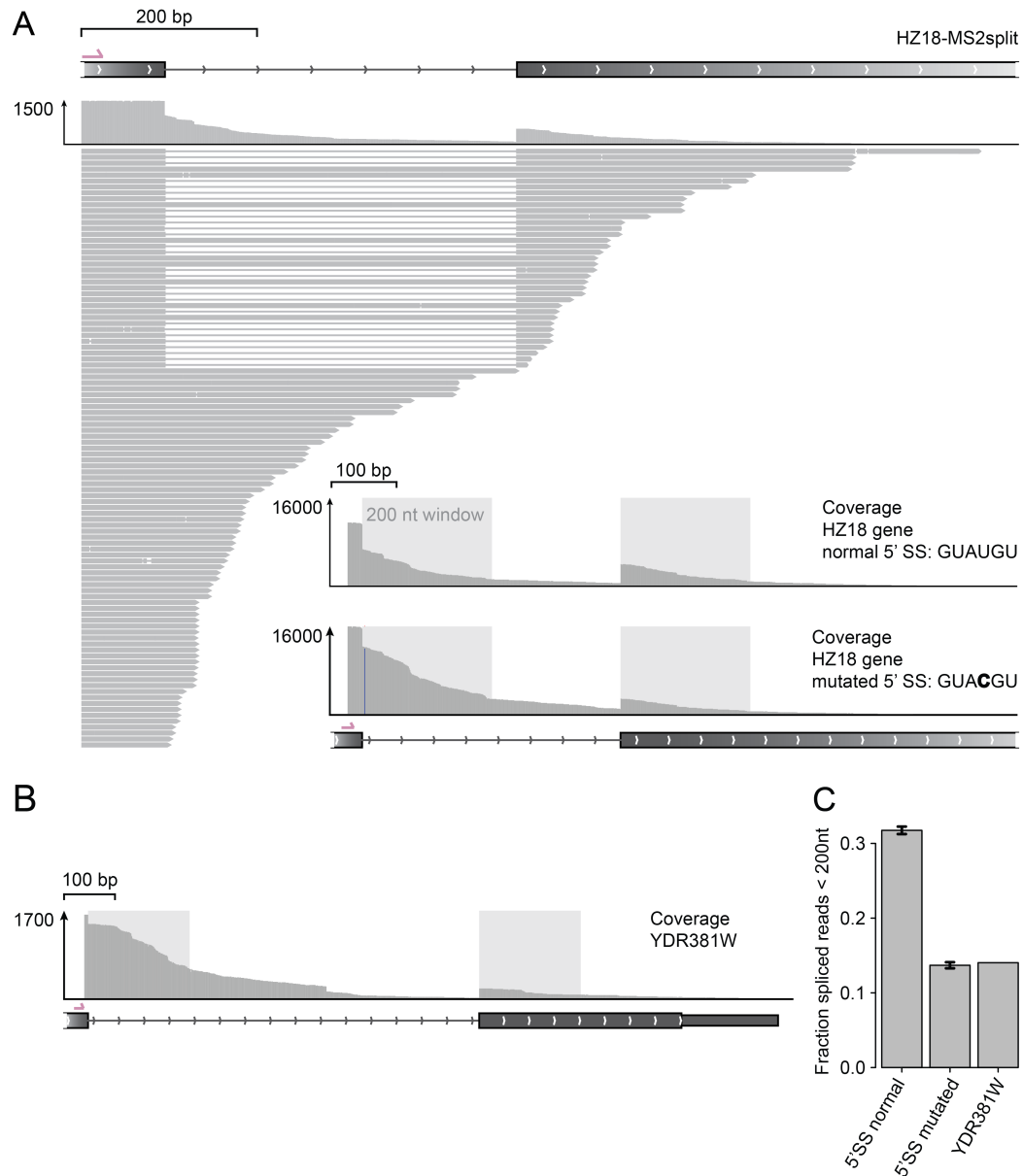


FIGURE 26 Early co-transcriptional splicing for splicing reporter gene. **A:** PacBio reads and coverage from double-stranded cDNA generated with a SMIT-like PCR for the HZ18-MS2split gene. The forward primer is located close to the end of the first exon (pink) and many co-transcriptional splicing events are visible. Inset: Coverage profile of the HZ18-MS2split gene with normal 5' SS (GUAUGU) or mutated 5' SS (U4C). Read counts over introns increase for the mutated transgene. Coverage within the 200 nt window (light grey) in the intron and 2nd exon is used to calculate the fraction of spliced transcripts among the pool of transcripts < 200 nt in **C**. **B:** Coverage profile of an endogenous yeast gene. The 200 nt window for splicing quantification is marked in light grey. **C:** Co-transcriptional splicing of the unmutated HZ18-MS2split gene is high within the first 200 nt with 32%. The 5' SS mutation reduces co-transcriptional splicing to the co-transcriptional splicing level of the assayed endogenous gene, which is later spliced co-transcriptionally (Figure 52, n=3 HZ18-MS2split, n=1 YDR381W, SD is shown).

Part III

DISCUSSION

CO-TRANSCRIPTIONAL SPLICING

6.1 GLOBAL CO-TRANSCRIPTIONAL SPLICING LEVELS, KINETICS AND TRANSCRIPTIONAL PAUSING

The consensus found in recent years is that pre-mRNA splicing commitment and catalysis happens predominantly co-transcriptional (summarized and discussed in [Brugiolo et al. 2013, Herzel and Neugebauer 2015] and Section 7.1). The *S. pombe* data on co-transcriptional splicing presented as part of this work agree with this (Section 3.1).

When during transcription pre-mRNA splicing takes place and how long it takes, has strong implications in understanding how transcription and splicing are linked and how alternative transcript isoforms can arise from alternative splicing. Albeit several labs approached the question, no consensus has been found yet and estimates range from seconds to minutes. In this study, I chose a novel strategy of paired-end and PacBio sequencing of nascent RNA to address this question.

Using the SMIT assay (Single Molecule Intron Tracking), which is described in detail in Chapter 5 and discussed in Section 7.3, I could determine the progression of co-transcriptional splicing depending on the distance to the intron end. Most introns assayed from *S. cerevisiae* are spliced to 50% saturation within 100 nt after full intron synthesis (Figure 25). This splicing distance is much shorter than previously estimated in yeast and implies that nascent RNA is immediately spliced after the transcript emerges from the RNA exit channel of Pol II (> 15 nt [Rasmussen and Lis 1993, Andrecka et al. 2008, Martinez-Rucobo et al. 2015]) within 3 sec assuming a constant elongation of 2 kb/min. However, the individual SMIT traces allow to reject the assumption that splicing and elongation rate are constant (Figure 24, Figure 52). Fitting an exponential model, which is described by one parameter, the fraction of splicing and elongation rate, does not explain the observed traces and thus no real estimate can be given for how long completion of pre-mRNA splicing takes after the intron has been synthesized. High-resolution estimates of transcription elongation profiles would be important to determine the time it takes for splicing. Nevertheless, the data implies that the fast estimates of 15 sec for pre-mRNA splicing are most accurate for the assayed genes in *S. cerevisiae*.

Recent estimates of transcription elongation rates in human cells showed that average elongation rates vary in a gene-specific manner from 0.5-2.4 kb/min [Jonkers et al. 2014]. Within transcription units transcription elongation is also far from uniform. This gene-specific variability in transcription elongation is most likely also true for yeast, for example a range in elongation rates from 0.8 kb/min-2 kb/min has been measured in a different genes [Mason and Struhl 2005, Zenklusen et al. 2008] and frequent Pol II pausing and Pol II backtacking

has been identified in *S. cerevisiae* genes [Churchman and Weissman 2011]. In higher eukaryotes promotor-proximal pausing is commonly observed [Jonkers and Lis 2015] and there is also evidence that Pol II slows down over exons, which then leaves more time for pre-mRNA splicing [Kwak et al. 2013, Jonkers et al. 2014]. Determining elongation rates individually for endogenous introns in comparison to the adjacent exons by linear regression from GRO-Seq data similar to the studies in mammalian cells [Jonkers et al. 2014] is challenging in yeast, because introns are very short in the range of 30-1000 nt compared to several kbs in mammals for example. A modification of SMIT or high-resolution (single gene) PacBio sequencing of nascent RNA could be a solution (Section 7.3). Nucleosomes have been shown to be preferentially positioned over exons [Schwartz et al. 2009, Tilgner et al. 2009] and could facilitate slowing down of Pol II [Churchman and Weissman 2011]. Complicating things, *S. pombe* nucleosome maps did not reveal such an exon-intron pattern [Moyle-Heyrman et al. 2013]. Therefore the question remains, how chromatin structure and regulation feed into the detected co-transcriptional splicing patterns.

Spliceosome assembly occurs sequentially ([Görnemann et al. 2005, Lacadie and Rosbash 2005, Tardiff et al. 2006], described in Chapter 1) with the possibility that the 5' SS is recognized by the U1 snRNP even before the intron is fully synthesized. Therefore, the time it takes to transcribe the full intron feeds also into the available time for co-transcriptional splicing and the first step in splicing could take place before the 3' SS is accessible. Interestingly, single intron genes with long introns are associated with high co-transcriptional splicing levels in *S. cerevisiae* [Carrillo Oesterreich et al. 2010] and high transcription and splicing factor recruitment (Figure 27). To illustrate this, genome-wide *S. cerevisiae* transcription and splicing factor chromatin immunoprecipitation data (ChIP-chip) from [Mayer et al. 2010, 2012, Meinel et al. 2013] were aligned relative to the mapped pause site in terminal exons of highly spliced *S. cerevisiae* genes. ChIP-chip data alignment for high expressed intronless genes and lowly spliced genes to an equidistant position towards the 3' end of the gene shows that high transcription and splicing factor recruitment cannot be detected in those genes. Thus this pattern is specific for this group of highly expressed, highly spliced intron-containing genes. The pausing site is roughly located in the middle of terminal exons [Carrillo Oesterreich et al. 2010] and thus further downstream of where the majority of co-transcriptional splicing takes place (Chapter 5). Therefore, this pause site might not be directly associated to the process co-transcriptional splicing, but could enable reformation of mRNPs. Pol II pausing in this case seems to resemble a transition between two stages of transcription elongation (one group of elongation factors binds before and one peaks at the pause site) and could be important for spliceosome disassembly and the transition to transcription termination with nascent transcript cleavage and polyadenylation for the final formation of an export-competent mRNP. A two-step 3' transition in transcription has been noted earlier for ribosomal protein genes in yeast, although no distinction was made in the analysis between intron-containing and intronless genes [Mayer et al. 2010]. Post-translational modifications of the Pol II CTD are generally associated with the distance from the transcription start site and the

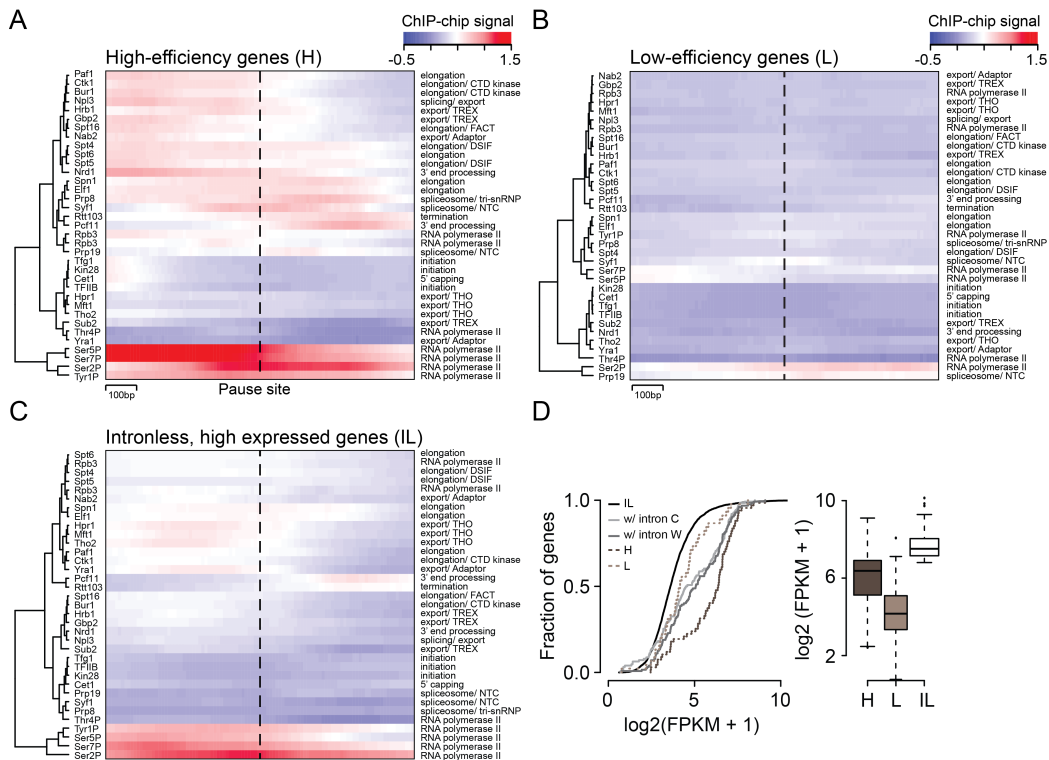


FIGURE 27 Transcription- and splicing factor distribution around the terminal exon pause site. 35 average ChIP-chip profiles for A: high efficiency genes (H, $n=73$), B: low efficiency genes (L, $n=47$) and C: highly expressed intronless genes ($IL > 800$ bp, $n=73$) are clustered according to their profiles 500bp up- and downstream of the pause site (discontinuity in A) or the corresponding position upstream of the polyA cleavage site in non-pausing low efficiency and intronless genes. D: Cumulative nascent transcript expression data for all intronless genes (black, $n=5,177$), intron-containing genes on Watson and Crick strand (two shades of grey, $n=126$ & $n=127$) and high and low efficiency genes (two shades of brown, $n=67$ & $n=45$) are shown. Intron-containing genes are much higher expressed than intronless genes and follow a bimodal expression pattern. Low efficiency genes are overall lower expressed than high efficiency genes. The three gene groups compared in A-C have very different nascent transcript expression patterns (box-plot in D).

phase of transcription, initiation, elongation or termination [Mayer et al. 2012]. A characteristic pattern is also observed around the pause site (Figure 27A). A decrease of the average signal of early CTD modifications (Serine 5 and 7 phosphorylation) around the pause site is observed and Serine 2 phosphorylation levels rise. To understand the interconnection between post-translational CTD modifications, transcription elongation factors binding patterns and co-transcriptional splicing around the terminal exon pause site in *S. cerevisiae* further experiments are necessary.

6.2 GENERAL AND *S. pombe* SPECIFIC INTRON SPLICING CHARACTERISTICS

S. cerevisiae as a single cell eukaryote serves as a strong model system to study conserved cellular processes [Forsburg 2005]. However, intron-containing genes are rare in *S. cerevisiae* and might represent a very specific class of genes with distinct properties, which helped them “survive” throughout evolution (explained in detail in Section 1.4). Therefore, I decided to include *S. pombe* as a model system to study co-transcriptional splicing. It is assumed that *S. pombe* contains a splicing machinery more closely reflecting the archetype of a spliceosome machinery than *S. cerevisiae*, shares higher similarity to the human splicing factors than *S. cerevisiae* [Kaeufer and Potashkin 2000] and contains > 1000 genes with multiple introns, making it a promising model system to study the order of intron removal.

The quantification of global co- and post-transcriptional splicing levels in *S. pombe* by nascent and mRNA-Seq of three different subcellular fractions showed that the majority of introns are spliced co-transcriptionally, albeit to a slightly lesser extent than in *S. cerevisiae* [Carrillo Oesterreich et al. 2010]. Single gene examples suggest that co-transcriptional splicing happens as close to intron ends as seen by SMIT in *S. cerevisiae*. (Figure 19, Figure 20, Figure 47, Figure 49).

Even though pre-mRNA splicing can happen as soon as the intron is synthesized, not all introns are spliced to the same extent. I observed frequent “not in order” intron splicing in transcripts spanning multiple introns, where first introns were most often not spliced (Section 4.3.3). In general, first introns and single intron genes were spliced less than internal and terminal introns, but also terminal introns were significantly less spliced compared to internal introns. This suggests that adjacent introns and splice sites positively influence pre-mRNA splicing of another intron. A similar observation was made in *D. melanogaster* before [Khodor et al. 2011].

The comparison of the *S. pombe* pre-mRNA splicing analysis to the published analyses of *S. cerevisiae*, *D. melanogaster*, mouse and human cells is given in Table 4. Many gene architecture features are significantly correlated with high or low co-transcriptional splicing in the evolutionary very distant species. This emphasizes the high conservation of pre-mRNA splicing and supports the use of the single cell organism *S. pombe* as model system for co-transcriptional splicing studies.

However, the main difference in gene architecture between *S. pombe* and the other species lies in intron length, which is also reflected in the opposite correlation found. In species with short introns it is believed that pairing between the splice sites takes place across an intron, if exons are separated by short (< 250nt) introns [Romfo et al. 2000, Fox-Walsh et al. 2005]. Thus, co-transcriptional splicing would be optimal for short introns. In species with generally longer introns, like mouse and human, longer introns tend to be better spliced. This could be associated with facilitated exon definition (splice site pairing across exons) [Roberson et al. 1990, Berget 1995]. *S. cerevisiae* is special in this case as it contains two classes of introns with different lengths (Figure 4A), but which can also be

GENE PROPERTY	IDENTIFIED ALSO IN
Low first intron splicing	<i>D. melanogaster</i> , mouse, human cells
Low single intron splicing	<i>D. melanogaster</i> , mouse
Internal introns better spliced than first and last	<i>D. melanogaster</i>
Optimal internal exon length 25-80 nt	50-500 nt mouse
Longer exons, lower co-transcriptional splicing	<i>S. cerevisiae</i> , <i>D. melanogaster</i> , mouse, human cells
High exonic GC-content, high co-transcriptional splicing	human cells
Modest correlation with splice site strength	human cells
Modest correlation with gene expression	<i>S. cerevisiae</i> , human cells
GENE PROPERTY	IDENTIFIED ONLY IN
Higher splicing for longer introns	<i>S. cerevisiae</i> , mouse, human cells
Lower co-transcriptional splicing closer to polyA site	mouse, human cells
GENE PROPERTY	IDENTIFIED ONLY IN
Low intronic GC-content, High co-transcriptional splicing; Optimal intron length 25-160 nt; Modest co-transcriptional splicing correlation with changes in expression	<i>S. pombe</i>

TABLE 4 Gene properties associated with pre-mRNA splicing: comparison with co-transcriptional splicing studies from human cells, mouse, *D. melanogaster* and *S. cerevisiae* [Tilgner et al. 2012, Khodor et al. 2012, 2011, Carrillo Oesterreich et al. 2010].

distinguished in their terminal exon lengths and gene expression (Figure 4A, Figure 27D, [Carrillo Oesterreich et al. 2010]).

The intron position is a strong determinant for co-transcriptional splicing in any species (*S. cerevisiae* excluded, mainly single-intron genes), albeit the correlation differs between *S. pombe* and *D. melanogaster* on the one side and mouse and human cells on the other side. Similar to *S. pombe*, *D. melanogaster*'s first and last introns are less well spliced than internal introns. Low splicing of first introns is also seen in mouse and human cells, but otherwise there is a decrease in intron splicing from the gene start towards the end of the gene.

Both, *D. melanogaster* and *S. pombe* have mainly short introns (median *D. melanogaster* 86 nt [Yu et al. 2002], median *S. pombe* 56 nt Figure 4B), which brings adjacent introns close together and could potentially serve a positive feedback interaction between adjacent splicing machineries. This might be less pronounced in genes/species that contain very long introns. It would be interesting to see how internal short introns are spliced with respect to each other in species like human or mouse, where only a minority of introns lie in the size range of *D. melanogaster* and *S. pombe* introns. In general, the high splicing of internal introns supports the observation in *S. cerevisiae* and *S. pombe* that co-transcriptional happens within a few nucleotides downstream of the intron end. There, one would not expect a strong correlation between lower co-transcriptional splicing levels and shorter distance to the gene end.

First and terminal intron splicing commitment and catalysis could be co-occurring with 5' capping or transcript cleavage and polyadenylation. Indeed, I observed significant lower co-transcriptional splicing for first intron shorter than 200 nt and terminal exons shorter than 300 nt, suggesting spatial constraints between transcription start site (TSS) and transcript cleavage site. Alternative TSS choice has been linked to first intron splicing in fly heads [Khodor et al. 2011] and is also prevalent in *S. pombe* [Li et al. 2015]. PacBio sequencing could be used to address this aspect further on single molecule level, where splicing levels and TSS can be directly linked. In addition to spacial constraints, roles for splicing factors close to the TSS have been identified, which might interfere with the efficient co-transcriptional splicing of first introns, but also link pre-mRNA splicing to transcription [Kwek et al. 2002, Damgaard et al. 2008, Kaida et al. 2010, Görnemann et al. 2005].

Similarly, interactions have been found between splicing and the transcription termination and cleavage machinery [Cooke et al. 1999, Dye and Proudfoot 1999]. Impaired spliceosome recruitment to a mutated model gene in human cells enhanced the accumulation of stalled Pol II downstream of the polyA site [Martins et al. 2011]. My analysis of *S. pombe* nascent RNA PacBio data showed a direct correlation between imperfect pre-mRNA splicing and transcript cleavage at the polyA site. Transcripts that were not cleaved at the polyA site were preferentially completely unspliced, suggesting that improper spliceosome assembly and incomplete intron splicing can effect transcript cleavage and might set the fate of those transcripts towards exosomal degradation. Interestingly, intronless genes showed a similar fraction of uncleaved transcripts like the unspliced uncleaved fraction of intron-containing genes. This suggests a positive role for co-

transcriptional splicing in enhancing transcription fidelity. Future experiments modulating this system, e. g. by splicing inhibition, are required to test this hypothesis.

Furthermore, first and terminal introns could be located within UTRs and thus serve a role in gene expression different from internal introns. About 35% of human 5' UTRs [Cenik et al. 2010] are annotated as intron-containing. The presence or absence of a 5' UTR intron can dictate the mechanism of mRNA export and some 3' UTR introns were found to target the mRNA for degradation by NMD [Bicknell et al. 2012]. How many introns are located within UTRs in *S. pombe* and if they show lower average pre-mRNA splicing patterns or are differentially regulated compared to introns within protein-coding regions remains to be seen. High levels of co-transcriptional splicing (this study and [Brugiolo et al. 2013]), splicing associated changes in transcription elongation [Alexander et al. 2010, Carrillo Oesterreich et al. 2010, Jonkers et al. 2014], functional links to transcription initiation [Cramer et al. 1997, Dujardin et al. 2013] and termination (this study and [Cooke et al. 1999, Dye and Proudfoot 1999]) and the association of 5' SS with active expression-associated chromatin marks [Bieberstein et al. 2012] form arguments that co-transcriptional splicing serves as a modulator in shaping gene expression profiles.

In line with this, *S. pombe* co- and post-transcriptional splicing are especially pronounced in genes with highest gene expression (Figure 13). I tested to which extent changes in mRNA expression are also reflected in changes of co-transcriptional splicing levels by treating *S. pombe* with caffeine. This drug is known to alter gene expression through a TORC1-associated pathway and results in downregulation of genes involved in cell growth and proliferation, e. g. genes with translation-associated functions, and upregulation of genes associated with nitrogen starvation, e. g. amino acid transport (Section 4.1 and [Rallis et al. 2013]). Of 1,512 genes with significant differences in mRNA expression upon 15 min caffeine treatment only 566 genes were intron-containing (37%). This is lower than the genome average of intron-containing protein-coding genes in *S. pombe* (48.5%) and probably reflects the general notion that intron-containing genes are underrepresented in rapidly regulated genes in response to stress in *S. pombe* [Jeffares et al. 2008]. Nevertheless, I detected a modest correlation between co-transcriptional intron splicing and mRNA expression changes (Figure 14), which is not apparent in the correlation of co-transcriptional splicing values with nascent RNA expression pointing to a role in RNA degradation or nuclear retention of unspliced transcripts to establish changes in mRNA expression. Further analysis of the data with regard to RNA stability and localization, and integration of existing mRNA half-life estimates [Lackner et al. 2007, Sun et al. 2012] are needed to provide a more detailed answer on which processes are involved in establishing changes in gene expression upon caffeine treatment.

Overall, the data presented in this thesis define the position of pre-mRNA splicing within the process of transcription and provide evidence for fast and efficient co-transcriptional splicing in *S. cerevisiae* and *S. pombe*, which is associated with high expressed genes in both organisms. Differences in *S. pombe* co-trans-

criptional splicing could be linked to gene architecture features, like intron position, GC-content and exon length.

STRATEGIES OF SEQUENCING NASCENT RNA

7.1 QUANTIFICATION OF CO-TRANSCRIPTIONAL SPLICING FROM RNA-SEQ DATA

In the recent years several papers were published on the global nature of co-transcriptional splicing (reviewed in [Brugiolo et al. 2013]). In order to decide for a computational strategy to quantify pre-mRNA splicing in *S. pombe* and to understand how different published co-transcriptional splicing calculation strategies impact splicing outcome, I applied several approaches in parallel to quantify co-transcriptional splicing from publicly available data [Khodor et al. 2012, Bhatt et al. 2012].

Using my analysis, I revealed an average of 60% co-transcriptional splicing for mouse liver (Section 3.1 and [Herzel and Neugebauer 2015]). This confirms the conclusion from [Khodor et al. 2012] that mouse co-transcriptional splicing is less efficient than in yeast [Carrillo Oesterreich et al. 2010], fruit fly [Khodor et al. 2011] and human cells [Tilgner et al. 2012], where the average intron or exon is spliced to ~75%. Three out of four strategies to quantify pre-mRNA splicing compared very well among each other and I decided to use the intron-centric “splicing per intron” measure for the *S. pombe* data, because it is most suitable for short introns, whose splicing is initiated by intron definition [Romfo et al. 2000, Fox-Walsh et al. 2005]. The average co-transcriptional splicing in *S. pombe* is lower than for other species (58%), but still very prevalent for most introns (Section 3.2).

The global co-transcriptional splicing value of the second dataset I analyzed (mouse macrophage chromatin-associated RNA [Bhatt et al. 2012]) was much higher with 90%, independent of the quantification strategy used (Figure 5D-F). This is also higher than reported previously using the same dataset [Bhatt et al. 2012]. This discrepancy could be explained by several things:

1. Bhatt et al.’s analysis only considers a group of genes with lowest nascent RNA expression. However, my and other studies show that pre-mRNA splicing correlates with gene expression (Section 4.1) [Cramer et al. 1997, Wilhelm et al. 2008].
2. Bhatt et al.’s publication does not state, which transcriptome mapping version has been used and whether a gene annotation was provided to aid junction detection. I provided the mm9 RefSeq gene annotation for mapping and used Tophat2 version 2.0.1.13. In a preliminary comparison mapping without gene annotation, indeed, identified less spliced junction reads and thus would lead to lower splicing estimates.
3. Bhatt et al.’s chromatin samples were not depleted for polyA+ RNA, which might artificially elevate co-transcriptional splicing levels. Polyadenylated

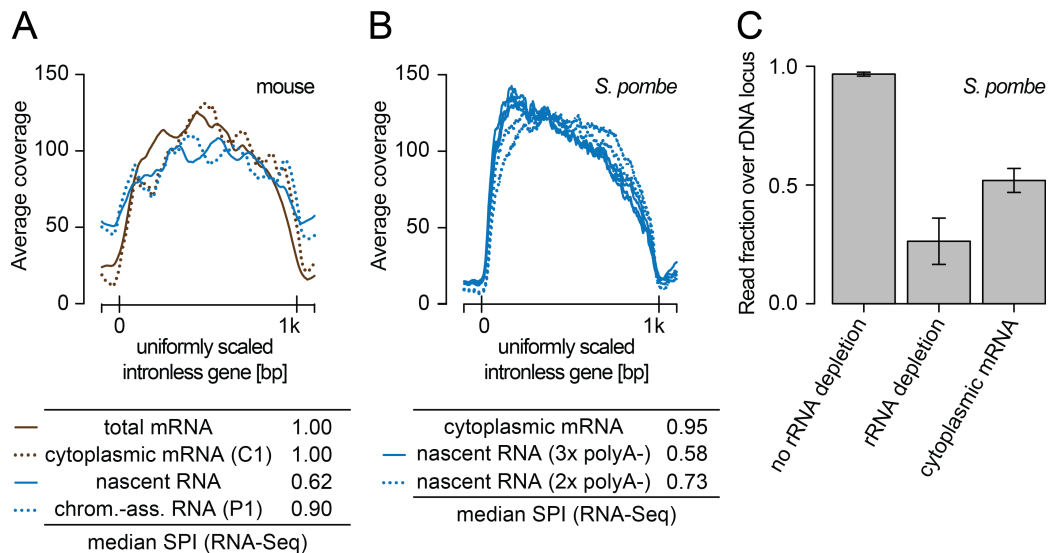


FIGURE 28 Influence of polyA+ RNA and rRNA depletion on splicing quantification. A: Average coverage profile over 149 high-expressed intronless genes ranging from 1-4 kb (coverage normalized for total sum $\cdot 10^5$) from mouse. Genes were scaled to 1 kb with additional 100bp added up- and downstream of the ORF. Nascent and mRNA-Seq data from [Khodor et al. 2012] and chromatin-associated RNA (not polyA+ RNA depleted) and cytoplasmic mRNA data from [Bhatt et al. 2012] were reanalyzed. The two chromatin-associated RNA samples differ in their global splicing values (table). B: Average coverage profile over 668 high-expressed intronless genes ranging from 1-4 kb. Genes were scaled to 1 kb with additional 100 bp added up- and downstream of the ORF. Data for three replicates of nascent RNA-Seq experiments are shown. One sample is 2x depleted for polyA+ RNA and one sample 3x. This makes a significant difference in global splicing quantification (table). C: rRNA removal is crucial for nascent RNA-Seq experiments to quantify pre-mRNA splicing. Barplot showing the average fraction of reads mapping to the rDNA locus in *S. pombe* on chrIII (n=3-6; SD is given).

RNA might be present in the sample (Figure 28A, Section 7.1.1). Thus, higher splicing levels than observed in the first dataset [Khodor et al. 2012] could be the consequence.

7.1.1 Removal of polyadenylated RNA

Polyadenylated RNA associated with chromatin might artificially elevate the quantification of co-transcriptional splicing levels by RNA-Seq. To cope with this point many nascent RNA-Seq protocols remove polyA+ RNA with oligo-dT coated cellulose [Carrillo Oesterreich et al. 2010, Khodor et al. 2011, 2012]. Furthermore, an enrichment of $\sim 40x$ for uncleaved nascent RNA relative to polyadenylated RNA was detected (Figure 8E, [Carrillo Oesterreich et al. 2010]). Dependent on the salt concentration, different stringencies in polyA tail removal can be achieved [Kojima et al. 2012]. Short polyA tracts within transcripts might be depleted together with polyA+ RNA, however, polyA tracts are rare in protein-coding genes [Koutmou et al. 2015] (1/200,000 for 15 consecutive As in exons or introns). In addition, my analysis did not reveal differences in the fraction of

A-tracts with varying length compared to the fraction of A-tracts in the sequencing data. For some polyA tracts longer than 15 nt, I observed reduced sequence coverage in comparison to the adjacent gene region. This might also be associated with the most common error in Illumina sequencing data, insertions and deletions after homopolymeric stretches [Quail et al. 2012], resulting in reduced mapping efficiency at those sites.

In addition to assessing removal of polyadenylated RNA by RT-qPCR, meta analysis of RNA-Seq coverage over intronless genes could be used. Nascent RNA sequencing coverage should show a 5' to 3' coverage bias, and mRNA sequencing coverage should form a block-like or 3' end biased profile. Figure 28A-B show average RNA-Seq coverage profiles over intronless genes from mouse and *S. pombe*. For both species, one sample with less or no removal of polyadenylated RNA (dotted blue lines) and one with removal (solid blue lines) is shown. The polyA-depleted sample has a high average signal downstream of the annotated polyA site, which is indicative for non-cleaved nascent transcripts. This signal is lower in the non-depleted sample, and the decrease in signal is steeper around the polyA site, making it more similar to mRNA. Similarly, in the *S. pombe* sample, a less pronounced 5' to 3' coverage bias is correlated with less depletion of polyadenylated RNA and higher quantified splicing levels in those samples.

7.1.2 Removal of ribosomal RNA

Using nascent and mRNA-Seq, I quantified intron splicing for 4,770 (90% of introns) and 2,282 (43% of introns) in 3 replicates for nascent RNA and cytoplasmic mRNA, respectively. I estimated a cutoff of 10 reads per junction for the data as described in Section A.2. Less than half of the number of introns could be quantified in cytoplasmic mRNA, even though all samples were sequenced with the same depth (Table 10), and the same cutoff was used. The analysis of the amount of rRNA still present in every sequencing sample shows that nascent RNA-Seq data only contain half the fraction of rRNA-associated reads than cytoplasmic mRNA, and thus less data depth is obtained for protein-coding genes in cytoplasmic mRNA (Figure 28C). Nascent RNA samples were depleted for rRNA using the commercially available RiboZero kit and mRNA was isolated with oligo-dT coated cellulose. The positive selection for mRNA is less efficient for rRNA removal than negative selection for rRNA species. However, this approach was chosen over rRNA depletion of the cytoplasmic fraction, which could also include RNA degradation intermediates, because I aimed to compare co-transcriptional intron splicing to splicing of fully-processed translatable fraction of mRNAs.

If rRNA is not removed prior to sequencing, 97% of sequencing data map to the rDNA locus (Figure 28C, first bar). In that case, quantification of pre-mRNA splicing would be impossible, but rRNA depletion might also deplete other RNAs, e. g. associated non-coding snoRNAs. This is important to consider when studying the biology of ncRNAs associated with chromatin, but influences less the quantification of pre-mRNA splicing in protein-coding genes.

7.2 CHROMATIN-ASSOCIATED TRANSCRIPTS FROM PACBIO SEQUENCING

Deep sequencing of many short fragments of transcripts is a valuable technique to study gene expression levels and pre-mRNA splicing levels. Software has been developed to identify and quantify various splice isoforms, e. g. MISO, spliceR, SplicingCompass and Cufflinks [Trapnell et al. 2012, Katz et al. 2010, Aschoff et al. 2013, Vitting-Seerup et al. 2014]. However, nascent RNA-Seq gene coverage often decays towards the 3' end of genes, complicating the isoform analysis in nascent RNA. Furthermore, a direct observation of how splicing of individual introns is connected to splicing of other introns in the same transcript or polyA site cleavage in nascent transcripts cannot be observed directly. The recent long-read sequencing development, Pacific Biosciences (PacBio) sequencing [Eid et al. 2009, Korlach et al. 2010], allows to detect those transcripts. Therefore, it is ideally suited to study multi-intronic transcripts from *S. pombe*, which fit the length profile of PacBio libraries. So far, PacBio sequencing has been mainly used for sequencing small microbial and viral genomes and to fill gaps in larger genomes [English et al. 2012], which could not be sequenced previously due to coverage biases of existing deep sequencing methods [Ross et al. 2013].

While establishing the nascent RNA PacBio library protocol, the first mRNA isoform data generated with PacBio sequencing were published [Au et al. 2013, Brinzevich et al. 2014, Kleinman et al. 2014, Larsen and Smith 2012, Schreiner et al. 2014, Sharon et al. 2013, Thomas et al. 2014, Tilgner et al. 2014, Treutlein et al. 2014, Zhang et al. 2014]. The transcriptome libraries are designed in similar ways, with oligo-dT reverse transcription priming, a template switching reverse transcriptase and a low-cycle PCR amplification to obtain sufficient amounts of cDNA for the final PacBio library preparation. If specific genes are targeted with PacBio sequencing, gene-specific forward primers are included in the low-cycle PCR.

This is very similar to the library preparation I designed, with the main difference that my protocol assays nascent RNA, which is not yet polyadenylated. In my protocol, the 3' end is therefore ligated to a DNA adaptor, which serves as primer binding site for reverse transcription and allows the generation of double-stranded cDNA from full-length nascent RNA. The ligation of the 3' end DNA adaptor is a universal tool, which not only provides the opportunity to sequence nascent RNA, but also to identify the Pol II position during transcription. With this knowledge, I could quantify the progression of co-transcriptional splicing during transcription (Chapter 5).

Due to the fact that PacBio sequencing is mainly applied to genome sequencing, no splicing-sensitive mapper has been developed to align long read transcriptome data. I compared two alignment tools, BLAT [Kent 2002] and GMAP [Wu and Watanabe 2005], which have been developed to align long cDNA sequences to the genome, e. g. from Sanger sequencing. One major advantage of GMAP for subsequent analysis is that the mapped sequences are provided as SAM-format, a format also used in RNA-Seq analysis.

In Section 4.3 multiple examples of sequenced transcripts derived from intronless and intron-containing genes are shown. The data presented in this section

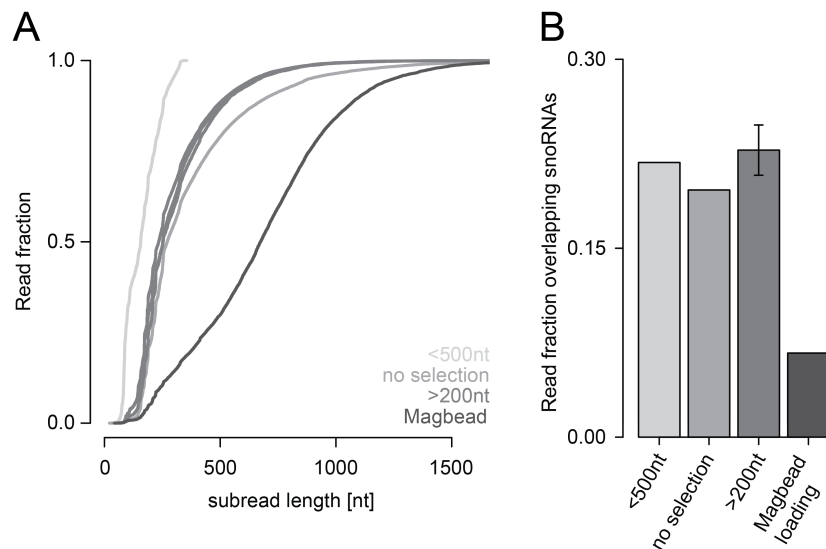


FIGURE 29 Read length considerations in PacBio sequencing. A: Cumulative distributions of PacBio subread lengths are shown for libraries prepared and sequenced in 3 different ways. Light grey: size selected RNA <500 nt; grey: conventional library (slightly darker grey: RNA >200 nt size selected); dark grey: conventional library with different SMRT cell loading (Magbead). B: Read fraction overlapping annotated snoRNAs does decrease with different SMRT cell loading, but not with RNA size selection ($n=1-3$, SD is given).

were all derived from pooled PacBio sequencing experiments shown in [Figure 18B](#).

7.2.1 snoRNAs in libraries with varying length spectrum

In the process of mapping and analysis, I noticed that the majority of reads originate from short ncRNAs, snoRNAs, which are not strongly enriched in this fraction, but get highly sequenced due to their short size (often <200 nt). In order to enhance future sequencing of protein-coding transcripts, I determined the snoRNA read fraction in PacBio libraries prepared with (a) size selection for short RNAs, (b) no size selection, (c) size selection against short RNAs and (d) a novel loading strategy of the SMRT cell (Magbead), which prefers DNA molecules longer than 800 nt ([Figure 29](#)). Selection against short RNAs did not reduce the fraction of snoRNAs. Furthermore, size selection for short RNAs did not markedly affect their abundance. However, the modified loading strongly reduces the amount of short RNAs and thus, also the fraction of snoRNA reads in the dataset. In this regard, the modified loading is a strong improvement.

A custom depletion approach, similar to what is commercially available (Epicentre/Life technologies) could also target abundant RNAs, e.g. rRNA or snoRNAs [[Rio et al. 2011](#)].

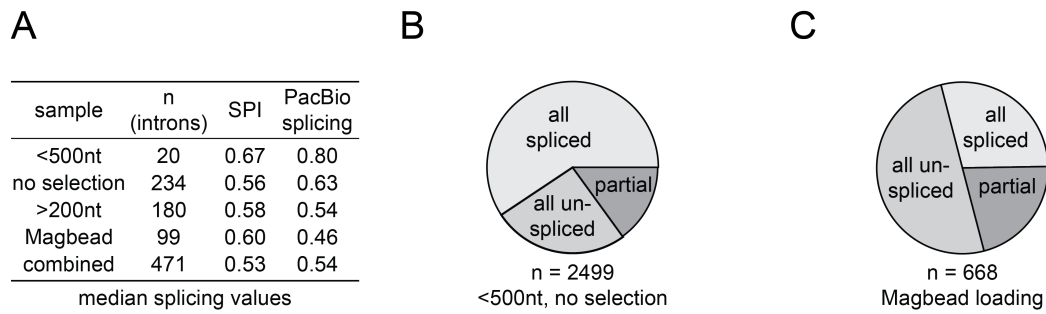


FIGURE 30 Average splicing levels in different PacBio libraries. A: The four different PacBio library types are compared for their median splicing level (spliced transcripts/all transcripts overlapping introns from single-intron containing genes, $n(\text{transcripts}) > 4$, not all (un)spliced). Median SPIs for the same intron groups are shown for comparison. The library with short transcripts overestimates co-transcriptional splicing, whereas the library loaded with Magbead loading underestimates it. B & C: Pie charts reflecting the fraction of all spliced, completely unspliced or partially spliced multi-intron-spanning transcripts. The fraction of all (un)spliced transcripts varies with the length distribution of sequenced PacBio libraries.

7.2.2 pre-mRNA splicing in libraries with varying length spectrum

The aspect that the PacBio sequencing with diffusion loading favors short transcripts and the one with Magbead loading longer transcripts, raises the question which of those two protocols - or maybe a combination of both - reflect the actual pool of protein-coding nascent RNAs best. This is especially important with respect to intron splicing analysis, as spliced and unspliced transcripts with the same 3' ends differ in length. In general, *S. pombe* introns are short with a tight distribution around the median length of 56 nt (Figure 4B), but the length difference increases with the number of spliced introns per transcript.

In single gene analyses, I observed a good correlation between nascent RNA-Seq SPIs and PacBio data, suggesting that the chosen approach represents the pool of protein-coding nascent RNAs well (Section 4.3). In addition, I calculated the fraction of spliced PacBio transcripts overlapping the intron of single intron genes and correlated this with intron splicing estimates from nascent RNA-Seq data for the different types of libraries (Section B.5, Figure 48). The number of genes included in the analysis varied depending on sequencing depth and library size distribution. The correlation is best for diffusion-loaded, non-size selected samples (0.33) and non-existing or low for the short transcript library and Magbead loading.

The comparison of median splicing values confirms that (a) splicing values from not size selected and diffusion loaded cDNA is most similar to the RNA-Seq splicing results and (b) sequencing only short transcripts overestimates co-transcriptional splicing and (c) Magbead loading underestimates splicing (Figure 30A). Consistent with that, the fraction of completely unspliced transcripts in the Magbead loaded sample is much higher than in diffusion loaded samples, being the dominant class of transcripts in this sample (Figure 30B-C). Considering that the global average in co-transcriptional intron splicing was determined to be 58% by

nascent RNA-Seq, the real amount of completely spliced and unspliced fraction lies in between the fractions shown in [Figure 30B-C](#). Independent of the method used, partially spliced transcripts are the least frequent class of transcripts. Sequencing nascent PacBio libraries with Magbead loading offers the potential for the analysis of how many genes produce long unspliced, non-terminated transcripts. My data showed a link between imperfect splicing and termination, but only for a small fraction of genes compared to the number of genes which had all spliced transcripts ([Section 4.3.4](#)). The widespread nature of this association between nascent RNA processing events will be an interesting aspect to follow up on.

7.3 SMIT AS GENERAL APPLICATION TO STUDY CO-TRANSCRIPTIONAL RNA PROCESSING

To obtain a quantitative image of co-transcriptional splicing kinetics, I developed an additional protocol for sequencing nascent RNA. With paired-end sequencing on the Illumina platform one can obtain millions of reads originating from single nascent RNA molecules. This harbors a huge potential for detecting 3' ends along genes at high resolution. The 3' ends resemble the position of Pol II during transcription and thus allow to determine, where the pre-mRNA is spliced during transcription. Therefore, the splicing status needs to be obtained. I achieved this by sequencing the 5' end of cDNA, which was generated in a defined PCR targeting the region in the exon just upstream of the intron of interest. This approach is called Single Molecule Intron Tracking (SMIT). With this technique at hand, I can detect single molecules due to a random barcode included in 3' end adaptor and track the intron presence in association with the Pol II position ([Section 5.1](#)). The only requirements for the assay I developed are detectable expression of the transcript, a first exon long enough to place a PCR primer inside and that the intron is significantly shorter than the average insert length of the sequencing machine (the gene with the longest intron of 1002 nt in *S. cerevisiae* cannot be assayed). I validated my assay using PacBio sequencing ([Figure 26](#) and [Figure 25C](#)). Similar to the insert length considerations discussed in the previous section ([Section 7.2](#)), also SMIT data show an insert length bias caused by (a) the clonal amplification on the flow cell [[Ross et al. 2013](#)], (b) better binding of short DNAs to the flowcell and (c) the SMIT PCR itself. PCR duplicates are removed during data processing ([Table 12](#)), but the enrichment of short inserts persists. Insert lengths follow an exponential distribution, which can be accounted for by normalization prior splicing quantification.

Initially, I developed the SMIT assay to quantify co-transcriptional splicing levels. However, it could very well be applied to other aspects of co-transcriptional RNA processing, e. g. to study Pol II pausing along genes. A previous study identified that Pol II pauses in short terminal exons of highly co-transcriptionally spliced *S. cerevisiae* genes [[Carrillo Oesterreich et al. 2010](#)]. This study was carried out by profiling nascent RNA with high-density tiling arrays, where a change in slope of the 5' to 3' intensity profile reflects a change in Pol II elongation behavior. I could detect higher Pol II density in those pausing genes in analysis of published Pol II

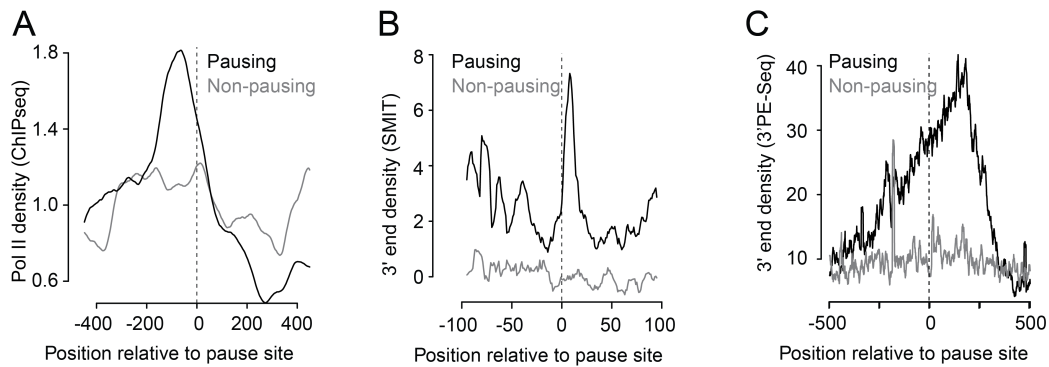


FIGURE 31 Detection of terminal exon pausing (TEP). A: Analysis of Pol II ChIP-seq data [Chathoth et al. 2014] for genes with (9 pausing genes, same as in B) and without (4 non-pausing genes, same as in B) pausing according to [Carrillo Oesterreich et al. 2010] (average coverage is shown). B: SMIT analysis of 9 genes with pausing and 4 genes without pausing reveals a strong peak of nascent RNA 3' ends immediately downstream of the pause site (average 3' end counts are shown). C: Nascent RNA 3' ends from 3' PE-Seq data for all identified 74 pausing genes and 48 non-pausing genes show a broad peak downstream of the pause site in pausing genes (total 3' end counts are shown).

ChIP-Seq data [Chathoth et al. 2014], as well as in SMIT 3' end data (first set of genes from Figure 23B, Figure 31A-B). Unfortunately, the underlying SMIT insert length distribution results in low coverage around the terminal exon pausing site. The assay could be adjusted by placing the forward primer not at the first exon end, but rather into the terminal exon, closer to the expected pause site. In a similar way to studying co-transcriptional splicing, also polyA site cleavage could be assessed by locating the forward primer close to the annotated polyA site.

Even though it is possible to account for the observed insert length bias in data post-processing, ideally, such a bias would already be reduced or eliminated during sample preparation. One strategy could be to replace the SMIT PCR with a linear enrichment step, e.g. by hybridization. A modification of the existing RNA CaptureSeq protocol [Mercer et al. 2014] could be developed, where 3' end ligated nascent RNA of intron-containing transcripts is enriched by hybridization against complementary oligonucleotides that are attached to magnetic beads. This process would be followed by Illumina sequencing. In a first experiment, I sequenced 3' end ligated nascent RNA from *S. cerevisiae* without enrichment by SMIT PCR or hybridization with paired-end sequencing (Figure 31C). I refer to this experiment as 3' PE-Seq. Despite some noise due to lower coverage, I could detect terminal exon pausing in a first meta analysis. It will now be interesting to see, if 3' end pausing profiles for single genes can be determined using hybridization-driven enrichment.

Interestingly, the 3' PE-Seq data is also useful to validate the original SMIT assay. Typically, sequencing libraries are generated using fragmented RNA or DNA of 100-300 nt [Nagalakshmi et al. 2008, Levin et al. 2010]. In the 3' PE-Seq experiment, 3' end ligated nascent RNA was fragmented first and then reverse transcribed using the SMIT DNA adaptor sequence as RT primer. Split reads carrying the 3' end adaptor, which originated from spliced RNA, must therefore

correspond to transcripts that were spliced in the window of 0-300nt downstream of the 3' SS. This is clearly reflected in the correlation of SMIT saturation values with 3' PE-Seq splicing values (Section C.1, Figure 53B). The SMIT saturation values from genes with fast splicing kinetics correlate highly with 3' PE-Seq splicing values, whereas the correlation is low for genes with SMIT saturation values reached far downstream of the 3' SS (Figure 53C). This is also reflected in the Pearson correlation of 0.54 between nascent RNA splicing values from tiling microarrays and the SMIT saturation values [Carrillo Oesterreich et al. 2010] (Figure 53A).

Overall, sequencing nascent RNA with SMIT and PacBio sequencing offers significant potential for the analysis of co-transcriptional splicing, transcription termination and RNA polymerase pausing. Although I developed the protocols for the analysis of co-transcriptional splicing, they could be well adapted for studying other aspects of gene expression.

CONCLUDING REMARKS

In this work, I developed and employed several deep sequencing strategies to study co-transcriptional splicing in the two distantly related yeast species, *S. cerevisiae* and *S. pombe*. In a first assessment of strategies to quantify co-transcriptional splicing levels, I made use of two published mouse nascent RNA-Seq datasets and compared existing approaches with respect to the quantification of co-transcriptional splicing in yeast. After establishing a nascent RNA-Seq workflow from chromatin for *S. pombe*, I quantified global co-transcriptional splicing levels in this species and found that the majority of introns are spliced co-transcriptionally in *S. pombe*, but with a slightly smaller global median than detected previously in *S. cerevisiae* [Carrillo Oesterreich et al. 2010]. In general, the combination of the two yeast species allowed to address multiple open questions about pre-mRNA splicing and its association to transcription and other RNA processing events.

Till this study, it was unclear, where and when co-transcriptional splicing takes place during transcription. With the development of a quantitative approach to measure the position of splicing relative to the end of an intron in endogenous genes, I detected the onset of pre-mRNA splicing immediately after the intron end emerges from the RNA exit channel of the transcribing Pol II. This is much closer to RNA polymerase II than previously estimated, albeit similar in distance to another RNA processing event, 5' capping of nascent RNA, which occurs 15-20 nt after transcription start [Rasmussen and Lis 1993, Martinez-Rucobo et al. 2015]. This new aspect needs to be considered in future analyses of pre-mRNA splicing, as it might have consequences how for example alternative splicing, e. g. exon skipping, is achieved.

Part IV

METHODS

METHODS

9.1 YEAST STRAINS

9.1.1 *S. pombe* strains

S. pombe chromatin fraction, RNA-Seq and PacBio experiments were performed using the *S. pombe* 972h- strain, for which the genome was first sequenced [Wood et al. 2002].

STRAIN	GENOTYPE
<i>S. pombe</i> Urs Leupold 972h-	972h-
<i>S. pombe</i> Urs Leupold 972h-	972h-
Sp-Prp5-WT	prp5::KanMx-prp5-WT
Sp-Prp5-APLD	prp5::KanMx-prp5-D303A
Sp-Prp5-DPAD	prp5::KanMx-prp5-L305A

TABLE 5 *S. pombe* strains9.1.2 *S. cerevisiae* strains

S. cerevisiae chromatin fraction and SMIT experiments were performed using the *S. cerevisiae* strain BY4741 (Mat a; his3D1; leu2Do; met15Do; ura3Do) obtained commercially (Euroscarf).

STRAIN	GENOTYPE
BY4741	MATa leu2Do met15Do ura3Do
468	MATa ade2 ade3 his3 leu2-3,112 trp1 URA3::Pgal1:pHZ18Split-MS2
475	MATa ade2 ade3 his3 leu2-3,112 trp1 URA3::Pgal1:pHZ18Split-MS2-GUAcGU

TABLE 6 *S. cerevisiae* strains

9.2 YEAST GROWTH

9.2.1 *S. pombe* growth

S. pombe were handled according to the Fission yeast handbook¹.

S. pombe cell cultures were grown in complete media (YES - 5 g Bacto Yeast extract, 225 mg Adenine, 225 mg Uracil, 225 mg L-Histidine, 225 mg L-Leucine, 225 mg L-Lysine, 3% D-Glucose in 1 L) at 30°C and 250 rpm.

For cell fractionation experiments cells were grown in 50 mL overnight to high densities and then diluted to an OD_{595 nm} of 0.2 in 1 L. Cells were harvested in exponential growth at an OD_{595 nm} of 0.5-1. For total RNA extraction dense 5-10 mL cell cultures were diluted to OD_{595 nm} 0.2 in 50 mL and harvested at an OD_{595 nm} of 0.5-1.

9.2.2 *S. cerevisiae* growth

S. cerevisiae BY4741 cell cultures were grown in complete media (YPD) at 30°C and 250 rpm. *S. cerevisiae* strains 468 and 475 were grown in YP supplemented with 1% Raffinose and 2% D-Galactose to induce gene expression of the assayed transgene under the GAL1 promotor.

For chromatin preparation experiments cells were grown in 50 mL overnight to high densities and then diluted to an OD_{595 nm} of 0.2 in 1 L. Cells were harvested in exponential growth at an OD_{595 nm} of 0.5-1. For total RNA extraction dense 5 – 10 mL cell cultures were diluted to OD_{595 nm} 0.2 in 50 mL and harvested at an OD_{595 nm} of 0.5-1.

9.3 CELL HARVEST

S. pombe or *S. cerevisiae* exponentially growing cell cultures at OD_{595 nm} of 0.5-1 in 500 mL-1 L were prepared.

9.3.1 Filtration

Cell filtration was performed as described by Churchman and Weissman [2012]. Pre-cut nitrocellulose membranes (90 mm diameter, 0.45 µm pore size) were placed on Microfiltration assembly (90-mm, ULTRA-WARE) and wetted with ice-cold PBS. The culture was rapidly filtered through. Yeast were then scraped from the filter, collected in 2 mL eppendorf tubes (6 per 500 mL culture) and immediately frozen in liquid nitrogen. Cell pellets were stored for nascent RNA extraction at -80°C.

¹ <http://www.biotwiki.org/foswiki/bin/view/Pombe/NurseLabManual>

9.3.2 Centrifugation

Cell filtration was performed as described by Carrillo Oesterreich et al. [2010]. All steps were done at 4°C and on ice. Cultures were poured into pre-cooled centrifugation buckets and centrifuged at 1,100 g for 5 min at 4°C. Cell pellets were washed once with 200 mL ice-cold PBS per liter culture. Subsequently, pellets were pooled and resuspended in 5 mL ice-cold PBS per liter culture. 1 mL aliquots were transferred into 2 mL eppendorf tubes and spun at 1,100 g for 5 min at 4°C. After supernatant removal samples were snap frozen in liquid nitrogen and stored at -80°C till further use.

9.4 CELL FRACTIONATION AND NASCENT RNA PREPARATION

9.4.1 Cell fractionation & RNA extraction

All steps were done at 4° and on ice. Yeast cell pellets harvested by filtration or centrifugation were used for this preparation and thawed on ice at 4°. Each cell aliquot was resuspended in 1 mL B1 buffer (Table 7) and the cell suspension was transferred into fresh a 2 mL eppendorf tube containing 1 mL Zirconia beads. Cells were vortexed in 5 (*S. pombe*) or 4 (*S. cerevisiae*) 1 min pulses at maximum speed. Inbetween cell-bead-suspension was kept on ice for 1 min. One 15 mL Falcon tube per sample was punctured and placed into a 50 mL Falcon tube carrying the tube lit with a circle cut with the approximate diameter of the smaller tube. The 15 mL Falcon tube was kept in place by wrapping Parafilm around at the 12 mL mark. The cell-bead-suspension was transferred to the punctured 15 mL Falcon tube and the 2 mL eppendorf tube was washed 3 times with B1 buffer to and the buffer was added to the cell suspension. Tubes were spun at 400 g for 5 min at 4°C. Without touching the cell pellet, 4x750 µL were transferred into 2 fresh 1.5 mL eppendorf tubes per sample and spun again at 400 g for 5 min at 4°C to eliminate unlysed cells. Cell lysate was transferred to fresh eppendorf tubes once more and then spun at 2,000 g for 15 min at 4°C. The supernatant was kept as cytoplasmic fraction. For protein analysis the fraction was snap frozen in liquid nitrogen and for RNA analysis immediately extracted with Phenol:Chloroform:IAA, pH 6.6 and the addition of 1% SDS.

The nuclear pellet was resuspended in 800 µL B1 buffer (Table 7) and the centrifugation was repeated. For analysis of nuclear mRNA nuclei were immediately lysed by addition of 1% SDS and RNA extracted with Phenol:Chloroform:IAA, pH 6.6. To proceed with the chromatin fractionation, the nuclear pellet was resuspended in B2 buffer (Table 8), vortexed for 5 sec and the sample was once more centrifuged at 20,000 g for 15 min and 4°C. After one 800 µL B2 (Table 8) was of the brownish dense pellet. The pellet was resuspended in 250-350 µL buffer P (Table 9). One volume of Phenol:Chloroform:IAA, pH 6.6 was added and the sample was incubated at 37°C and 1,150 rpm for one hour. To separate aqueous and organic phase from each other samples were spun at room temperature at maximum speed. ~80% of the aqueous phase were transferred to a 1.5 mL eppendorf tube.

9.4.2 Ethanol precipitation

1/10th of the sample volume of 3 M Sodiumacetate, pH 5.3, was added to the samples. At least 2.5 volumes of ice-cold 100% Ethanol were added to precipitate the DNA and RNA in the sample. Samples were placed at -80°C for at least 30 min.

Precipitates were collected by centrifugation at 20,000 g for at least 30 min and 4°C . The supernatant was discarded and the pellet was washed once with $\sim 300\ \mu\text{L}$ 80% ice-cold Ethanol. Centrifugation at 20,000 g for at least 5 min and 4°C followed. All supernatant was discarded by pipetting and the pellet was dried for ~ 6 min at 37°C . Samples were resuspended for further use in DEPC-water and stored at -80°C .

9.4.3 DNase treatment

To 80 μL RNA of a nascent RNA preparation 10 μL 10x TurboDNase buffer and 10 μL TurboDNase (2 U μL , Life technologies) were added. The sample was incubated for 30 min at 37°C . The digested RNA was purified with the RNA Clean & Concentrator-5 kit from Zymoresearch² and eluted with 80 μL DEPC-water. The TurboDNase treatment and subsequent column purification were repeated once more. Elution was done in 200 μL DEPC-water.

9.4.4 Removal of polyA+ RNA

For most parts the oligo-dT coated cellulose³ was used (SMIT, *S. pombe* and *S. cerevisiae* PacBio and RNA-Seq experiments). To 250 μL RNA solution 250 μL 2x Binding buffer were added. The solution was transferred to the cellulose-containing tubes, resuspended by pipetting and slightly vortexing. Denaturation of the solution at 75°C for 5 min followed. Afterwards, the samples were rotated at room temperature for 60 min. Cellulose and solution were transferred onto a filter cartridge in a 2 mL tube and for 5 min at 4,000 g. The flow through was kept and 2 more depletions of polyA+ RNA followed. Preceding the 60 min incubation the RNA was always denatured at 75°C for 5 min. For the *S. cerevisiae* 3'PE-Seq and SpPrp5 PacBio libraries oligo-dT coated magnetic beads⁴ were used. Also three rounds of enrichment were done with this kit and the polyA- fraction was always kept.

For both polyA+ RNA depletion protocols RNA was precipitated with Ethanol after (Section 9.4.2).

9.4.5 Qualitative and quantitative analysis of nucleic acids

RNA and DNA samples were analyzed by agarose (1-1.5%) or TBE-Urea polyacrylamide (10 or 15%, Invitrogen) gel electrophoresis. 500 ng-1 μg were loaded

² <http://www.zymoresearch.de>

³ MicroPoly(A)Purist kit, Life technologies

⁴ Dynabeads mRNA DIRECT Micro Purification Kit, Life technologies

per lane. Prior analysis RNA samples were denatured in Novex TBE-Urea Sample Buffer (2X, Life technologies) at 65°C for 5 min and immediately cooled on ice for at least 1 min. 0.5 µL O'GeneRuler 1 kb Plus DNA Ladder and O'GeneRuler Ultra Low Range DNA Ladder (Life technologies) were used for nucleotide analysis per gel.

DNA and RNA concentrations were determined by UV/Vis spectroscopy with the Nanodrop2000 (ThermoScientific) or fluorometric measurements with the Qubit dsDNA BR Assay or the RNA BR Assay (Life technologies).

Prior sequencing and RNA-Seq library preparation RNA quality was assessed with the Bioanalyser instrument and the RNA Nano kit (Agilent).

9.4.6 Removal of rRNA

Highly abundant rRNA was removed from the sample with the Ribo-Zero Gold rRNA Removal Kit (Yeast) (Epicentre/Illumina). Other available kits (Terminator 5'-Phosphate-Dependent Exonuclease (Epicentre/Illumina) and RiboMinus (Life technologies)) and enzymes were also tested, but are less efficient in rRNA removal in *S. pombe* (validated by RT-qPCR and Bioanalyzer chromatograms).

9.4.7 Buffers

COMPONENT	FINAL CONCENTRATION
HEPES, pH 8.0	20 mM
KCl	60 mM
NaCl	15 mM
MgCl ₂	5 mM
CaCl ₂	1 mM
Triton X-100	0.8%
Sucrose	0.25 M
Spermidine	2.5 mM
Spermine	0.5 mM
DTT	1 mM
PMSF	0.2 mM

TABLE 7 Buffer 1

COMPONENT	FINAL CONCENTRATION
HEPES, pH 7.6	20 mM
NaCl	450 mM
MgCl ₂	7.5 mM
EDTA	20 mM
Glycerol	10%
NP-40	1%
Urea	2 M
Sucrose	0.5 M
DTT	1 mM
PMSF	0.2 mM

TABLE 8 Buffer 2, pH 7.6

COMPONENT	FINAL CONCENTRATION
Sodium acetate	50 mM
NaCl	50 mM
SDS	1%

TABLE 9 Buffer P, pH 5.0

9.5 ISOLATION OF TOTAL RNA

Total RNA from *S. cerevisiae* and *S. pombe* (often used for testing different steps in protocol development or to isolate polyA- RNA, e. g. in SMIT) was extracted with Phenol:Chloroform:IAA, 25:24:1, pH 6.6 using the RiboPure RNA Purification Kit, yeast (Life technologies).

9.6 RT-(Q)PCR

9.6.1 Reverse transcription (RT)

Reverse transcription of RNA was done using SuperScript III reverse transcriptase (Invitrogen). The protocol recommended for the enzyme was used in either 10 or 20 μ L reactions. Depending on the application different amounts of RNA has been used:

- 60 ng for enrichment analysis of nascent RNA, splicing and abundance quantification of snoRNAs
- ~600 ng 3' end ligated RNA
- 1 μ g for circular RNA analysis and RNaseR sensitivity

Reverse transcription primers were used in the following final concentrations:

- Random hexamers (Roche, 1 μ L/20 μ L RT reaction)
- 0.1 μ M gene-specific primers, e. g. post polyA site primer
- 5 μ M oligo-dT primer
- 25 nM SMIT RT primer

RNA, dNTPs and RT primer were denatured at 65°C for 5 min and immediately cooled on ice for at least 1 min. Afterwards RT buffer, 0.1 M DTT and RNaseOUT were added in the respective amounts. Superscript III enzyme was added last (or water for -RT controls). Samples with random hexamer priming were incubated at room temperature for 5 min. RT samples were incubated at 55°C for 30 min and the enzyme was inactivated with a 70°C incubation for 15 min. cDNA samples were diluted 1:10 for further use or 1:2-1:5 in SMIT.

9.6.2 qPCR

Sybrgreen (Life technologies) qPCR reaction mix was used for qPCR assays. In each well of a 96-well plate 5 μ L Sybrgreen qPCR reaction mix, 3 μ L primer solution and 2 μ L 1:10 diluted cDNA solution were added. Each sample was assayed in technical triplicates and at least two no template controls were performed. Each run was done as recommended by the manufacturer. Optimal primer concentrations were determined in a test run for 3 primer concentrations (250 nM (final 83 nM), 500 nM (final 167 nM), 1 μ M (final 333 nM)) and four different cDNA

dilutions for 1:10, 1:100, 1:500 and 1:1000. Primer efficiencies in a range of 95-105% were accepted for further use and the determined cDNA levels were adjusted accordingly.

9.6.3 PCR

For all PCRs, except for PacBio libraries, Phusion DNA polymerase (NEB) was used with recommended reaction settings and reagent concentrations. PCR reactions were carried out in 10 or 20 μ L reaction volumes with 1-3 μ L 1:10 diluted cDNA template. 1 μ M PCR primer solutions were used (final 100 nM). The DNA synthesis time was set to 1.5 min for full-length products up to several kb and to 5 sec for short cDNAs.

9.7 DNA PURIFICATION

DNA amplified by PCR was purified with the MinElute PCR Purification Kit (Qiagen) after SMIT-PCRs, with AMPure beads (Beckmann Coulter) for PacBio sequencing or DNA precipitation with Ethanol (Section 9.4.2).

9.8 PROTEIN ANALYSIS

Western blot of different samples taken during *S. pombe* cell fractionation was performed to assay enrichment of proteins characteristic for chromatin. The Bradford Protein Assay (Bio-Rad, Assay protocol⁵) was used to determine protein concentrations. BSA standard curves were prepared in triplicates. Samples were adjusted to neutral pH for protein measurement. \sim 3 μ g of protein were loaded per lane in western blot analysis. Antibodies against two nuclear proteins, the largest Pol II subunit Rpb1 (8WG16), Histone H3 (ab1791, Abcam), and two cytoplasmic proteins, GAPDH (Novus Biologicals, NB300-221) and ribosomal protein L5 (Santa Cruz, sc-103865), were used. Coomassie gel staining and mass spectrometry was performed as described elsewhere [Shevchenko et al. 2006, 2008] at the mass spectrometry facility of the MPI-CBG Dresden.

9.9 RNA-SEQ

For RNA-Seq of different cellular fractions of *S. pombe* RNA samples were submitted to the Yale Center for Genome Analysis (YCGA). PolyA+ RNA depleted, rRNA depleted nascent RNA, cytoplasmic polyA+ RNA and nuclear polyA+ RNA was analyzed by RNA-Seq. PolyA+ RNA was prepared with protocols highlighted in Section 9.4.4. Random hexamer primed libraries were prepared with standard Illumina library protocols⁶. Single-end sequencing with 76 bp read length was done. Samples were sequenced in triplicates (nascent RNA and cytoplasmic mRNA) or duplicates (nuclear mRNA).

⁵ http://www.bio-rad.com/LifeScience/pdf/Bulletin_9004.pdf

⁶ <http://medicine.yale.edu/keck/ycga/sequencing/Illumina/protocols.aspx>

Paired-end sequencing for SMIT libraries was done either at the Sequencing Core Facility at the MPI-MG, Berlin (initial dataset, read length 2x150bp) or at the YCGA (size selection datasets and expanded gene set, read length 2x76bp). Library quality was assessed with Qubit and Bioanalyzer (Section 9.4.5).

9.10 3' END LIGATION

The library preparations for SMIT and PacBio sequencing (Section 9.11 and Section 9.12) require both 3' end ligation of a DNA adaptor to label the nascent RNA 3' end and to obtain a universal sequence for reverse transcription. 600 ng RNA and 0.5 μ L (50 pmol) of the 100 μ M threeend DNA adaptor (Table 15) were combined and DEPC-water was added to a final volume of 6 μ L. After 65°C for 5 min and at least 1 min on ice, 2 μ L 10x ligation buffer (final 50 mM Tris-HCl, 10 mM MgCl₂, 1 mM DTT, pH 7.5), 10 μ L 50% PEG 8000 (final 25%) and 1 μ L RNase-OUT were added. Samples were mixed thoroughly and after the addition of the T₄ RNA ligase II, truncated K227Q (200 U/ μ L, NEB) samples were incubated for 10 hours at 16°C. A minus enzyme control was always included in all experiments. After incubation, the 20 μ L reaction was diluted with 80 μ L DEPC-water to ease pipetting and RNA purification. RNA column purification was done to remove unligated adaptor and enzyme with the RNA Clean & Concentrator-5 kit from Zymoresearch⁷.

The amount of ligated RNA was determined by quantification of ligated and unligated product after denaturing TBE-Urea polyacrylamide gel electrophoresis with Fiji⁸ for *in vitro* transcribed RNAs (Section 9.13).

9.11 PACBIO NASCENT RNA LIBRARIES

In order to obtain RNA processing information of full length nascent RNAs PacBio sequencing libraries have been prepared. The design and different primer binding sites and be found in Figure 46. DNase treated, polyA+ RNA depleted, rRNA depleted nascent RNA from yeast was prepared (Section 9.4). Nascent RNA was 3' end ligated to the DNA adaptor (Section 9.10). The SMARTer PCR cDNA Synthesis Kit (Clontech) was used for reverse transcription with a slight modification. Instead of the included 3' SMART CDS Primer II A, a custom primer reverse-complement to the 3' end adaptor was used (Table 28). Maximal 1 μ g of RNA was used per reaction. The subsequent low cycle PCR was carried out with the Advantage 2 PCR Kit. For defining the optimal PCR cycle number, a 100 μ L PCR (maximum cDNA input ~100 ng RNA) was performed and split into 5 μ L after 9 cycle numbers. With the small aliquots the PCR was continued to different cycle number ranging from 12 to 25 in total and analyzed by agarose gel electrophoresis. Usually, ~13 cycles were optimal. PCR reactions were repeated, pooled and quantified using the Qubit dsDNA BR assay (Section 9.4.5). ~1 μ g

⁷ <http://www.zymoresearch.de>

⁸ <http://fiji.sc/Fiji>

of double-stranded cDNA was submitted for further PacBio library preparation and sequencing to the YCGA with standard protocols from Pacific Biosciences⁹.

9.12 SINGLE MOLECULE INTRON TRACKING (SMIT) LIBRARIES

The Single Molecule Intron Tracking assay allowed to profile co-transcriptional state of single transcripts and the associated distance of the transcript end relative to the 3' SS. The data were generated through paired-end sequencing. For this a particular library design had been developed (Figure 50). DNase treated, polyA⁺ RNA depleted, rRNA depleted nascent RNA from yeast was prepared (Section 9.4). Nascent RNA was 3' end ligated to the DNA adaptor (Section 9.10) and reverse transcribed into cDNA with a primer complementary to the DNA adaptor (Section 9.6.1, Table 15). The resulting cDNA solution was diluted by 1:2-1:5 and used as template for gene-specific 10 μ L SMIT-PCRs (Section 9.6.3, Table 15, Table 16). The forward primer was designed to bind to a gene-specific sequence, whereas the reverse primer targets the cDNA 3' end with the adaptor sequence. The individual PCRs were pooled and PCR purified (Section 9.7). The pooled cDNA was template for a second low-cycle PCR, which was necessary to attach final sequencing adaptors. This last step was repeated multiple times to obtain enough cDNA for sequencing. Another PCR purification was done (Section 9.7). All PCRs needed to be carefully adjusted in their numbers of cycles. Therefore groups of 20-30 genes with similar expression patterns [Gu et al. 2015] were prepared in one step and tested for the optimal combination of PCR cycles, e. g. 5, 10 or 15 cycles in the 1st PCR and 5,10 or 15 cycles in the 2nd PCR. Most samples were amplified with a combination of 10 and 12-14 cycles. Prior sequencing the double-stranded DNA ranging from ~150bp (no insert product) to >2 kb was purified by gel electrophoresis. Paired-end sequencing of DNA >200bp was done after quality control steps performed in the sequencing facility (Section 9.9).

The final size selection reduced the fraction of DNA inserts <100bp. To account for this SMIT libraries targeting the same genes, but with size-selected RNA were prepared. As before nascent RNA was 3' end ligated, but then precipitated in ethanol to ensure that all short RNAs are present in the pool. About 15 μ g of RNA were size selected from 25 to 250 nt from denaturing 10% TBE-Urea polyacrylamide gel electrophoresis as previously described [Churchman and Weissman 2012] and Section 9.4.5. This high amount of input RNA was required due to the small fraction of short RNAs and the limited recovery during the extraction. Samples were precipitated in ethanol and salt and samples originating from 4 lanes (4 μ g) were pooled and precipitated again. The amount of RNA was assessed by UV/Vis spectroscopy (Section 9.4.5) and 600 ng were used for RT (Section 9.6.1, Table 15). The resulting cDNA solution was diluted 1:2 and used for SMIT-PCRs, similarly in cycle number as described above. The DNA synthesis time in the PCR was shortened to 5 sec. The cycle number evaluation and final samples analysis were done by 8% TBE-polyacrylamide gel electrophoresis. AM-

9 www.pacificbiosciences.com

Pure bead PCR purification (0.8 μ L bead solution/1 μ L PCR reaction) was used to eliminate free nucleotides and primer sequences prior paired-end sequencing.

9.13 *in vitro* TRANSCRIPTION (IVT)

Plasmid or genomic DNA were amplified and linearized prior to IVT with PCR primers carrying the T7 or SP6 promotor (Table 26). The PCR product was purified with the QIAquick PCR Purification Kit (Qiagen) and IVT was carried out according to the instructions of the MEGAshortscript Kit (Life technologies). The RNA product was analyzed by 10 or 15% denaturing TBE-Urea polyacrylamide gel electrophoresis and used for experiments on the optimization of the 3' end ligation (Section 9.10).

9.14 MAPPING OF RNA-SEQ DATA

In order to determine global pre-mRNA splicing and gene expression values in *S. pombe* RNA-Seq data (Fastq files) were assessed for quality with the FASTQC toolkit¹⁰ and then quality filtered (FASTX Toolkit version 0.0.13¹¹) and mapped to the genome with Tophat2 (version 2.0.12) [Kim et al. 2013] using the following settings: fastq_quality_filter -Q 33 -q 20 -p 90; tophat2 -p 5 -i 30 -I 900 -g 1 -N 2 -G <Spombe_EF2> -segment-length 25 -library-type fr-firststrand -min-anchor-length 8 -splice-mismatches 0 -min-coverage-intron 30 -max-coverage-intron 900 -min-segment-intron 30 -max-segment-intron 900.

For indexing, file conversion and analysis samtools version 1.1 [Li et al. 2009] and bedtools version 2.20.1 [Quinlan and Hall 2010] were used. Sequencing data were visualized using the IGV genome browser [Robinson et al. 2011, Thorvaldsdóttir et al. 2013].

To identify potential circular RNAs in the *S. pombe* transcriptome an alternative RNA-Seq mapper was used (segemehl version 0.1.7 [Hoffmann et al. 2014]): segemehl.x -d <Spombe_EF2.fasta> -i segemehl/Spombe.idx -t 4 -S -A 99.

Coverage data (bedgraph-format) were obtained using the samtools view and depth function to first extract mapped reads per DNA strand and then converting them into coverage data (wig-format). Wig-files were converted with awk into the bedgraph-format. To facilitate visualization bedgraph files were converted in addition to the binary tdf-format with igvtools¹².

9.15 QUANTIFICATION OF INTRON AND EXON SPLICING LEVELS

Different ways to determine pre-mRNA splicing levels per intron or exon from nascent RNA-Seq data were applied previously [Khodor et al. 2011, Ameer et al. 2011, Tilgner et al. 2012] and compared with each other using published mouse nascent RNA-Seq data. The analysis is described in detail in Herzel and Neugebauer [2015].

¹⁰ <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

¹¹ www.hannonlab.cshl.edu/fastx_toolkit

¹² <https://github.com/igvteam/igv/>

For *S. pombe* and *S. cerevisiae* intron splicing calculation a similar workflow was applied. Junction reads originating from spliced (split, cigar contains “N”, but no other elements (I,D,S etc.)) and unspliced (full-length, cigar “read lengthM”) transcripts were extracted from all mapped reads using the following shell commands and tools (samtools 1.1 and bedtools2 2.20.1). Overlaps with annotated 5’ SS and 3’ SS junctions were searched for split and unsplit reads separately with bedtools intersect. An overlap of at least three nucleotides on each side of the junction was required. Output from junction read isolation (spliced and unspliced) were read into a list in R version 3.1.2¹³ and spliced and unspliced reads per junction were summed up per junction. The fraction of splicing was calculated for each splice site (5’ SS and 3’ SS) and the combination of both in the following way:

$$\text{SPI}_{5' \text{ or } 3'} = \frac{S}{u_{5' \text{ or } 3'} + S} \text{ and } \text{SPI} = \frac{S}{\frac{u_{5'} + u_{3'}}{2} + S},$$

with S corresponding to the count of spliced reads and U corresponding to the count of unspliced reads at the respective junction.

This resulted in a splicing score ranging from 0 to 1 with 1 being 100% spliced. A cutoff of at least 10 reads per junction was applied.

9.16 ANALYSIS OF PRE-MRNA SPLICING CHARACTERISTICS

9.16.1 GO term analysis

Gene Ontology (GO) analysis was performed using the R package topGO version 2.18.0 [Alexa et al. 2006]. Enrichment of GO-terms was tested with the “weight” algorithm. The p-value was determined with Fisher’s exact test. GO term annotations were retrieved from PomBase¹⁴. The fraction of genes/proteins with the associated GO term within the data set and of the total number of genes/proteins with this GO term was determined. The 3-6 most enriched terms for a certain classification (biological process, cellular component or molecular function) representative for the analysis are depicted in the figures.

9.16.2 Gene expression analysis

To determine gene expression values between replicates and different samples cufflinks version 2.2.1 and cuffdiff were used with the following settings: cufflinks -p 20 -G <Spombe_EF2> -b <Spombe_EF2.fasta>; cuffdiff -frag-bias-correct <Spombe_EF2.fasta> -num-threads 20 -library-type fr-firststrand -library-norm-method geometric <Spombe_EF2.gtf>.

FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) values were required to be greater than the difference between high and low confidence boundary [Nagaraj et al. 2011] and assigned with a cufflinks flag “OK”. The

¹³ <http://cran.r-project.org/>

¹⁴ <http://www.pombase.org/>

expression values calculated with cufflinks were used for Pearson correlation between replicates. The cuffdiff results taking replicates into account were used for differential expression analysis (Section 4.1) and for the correlation to SPIs.

9.16.3 Gene architecture analysis

Intron, exon and gene sequences and coordinates were extracted from the *S. pombe* EF2 genome sequence and annotation file from iGenomes (Illumina sequencing)¹⁵. The bedtools software suite and R Studio version 0.98.1103 were used to search for feature overlaps, statistical analysis and data visualization. The following packages were implemented into R for the analysis: caTools, Hmisc, ggplot2, gplots, venneuler, vioplot, stringr, hydroGOF, plyr, reshape, boot.

9.17 PACBIO DATA PROCESSING & MAPPING

PacBio transcriptome data were obtained in Fastq-format, assessed for their quality (Section 9.14) and filtered and trimmed for the 3' end DNA adaptor and downstream Clontech adaptor sequences (Figure 46) with cutadapt [Martin 2014] (-a CTGTAGGCACCATCAAT -n 1 -O 17 -e 0.1 -match-read-wildcards -discard-untrimmed -m 15). Trimmed sequences longer than 15 nt and with a maximum error of 10% were kept for further processing. Also the reverse-complement adaptor was trimmed (-g ATTGATGGTGCCTACAG). PacBio sequencing was not strand-specific, but the adaptor sequence allowed to retrieve strand information. A custom shell script generated the reverse-complement sequence for reads originally containing the reversed adaptor sequence. Trimmed fastq-files were concatenated and the remaining 5 nt random 3' barcode was removed using the FASTX toolkit (fastx_trimmer -Q 33 -t 5, Section 9.14). Another 5' adaptor removal step ensured that only reads were included in the downstream mapping, which carried 5' and 3' end adaptors (cutadapt -g AAGCAGTGGTATCAACGCA-GAGTACATGGG -n 3 -O 30 -e 0.1 -match-read-wildcards -discard-untrimmed -m 10). Processed fastq-data were mapped to the *S. pombe* or *S. cerevisiae* genome using gmap (-d <genome_index> -min-intronlength=30 -intronlength=850 -local-splicedist=850 -totallength=850 -trimendexons=0 -microexon-spliceprob=0.5 -direction=auto -find-shifted-canonical -allow-close-indels=2 -npaths=1 -nofails -fails-as-input -mapboth -A -format=samse) [Wu and Watanabe 2005]. The output contained mapped and unmapped reads, which were separated subsequently. Anchoring of mapped transcripts within +/-200 nt of annotated transcription start sites was required and transcripts ending within +/-100 nt of an annotated polyA site and short polyA tails (> 4 nt) were removed from the dataset. Visualization was done with the IGV genome browser (Section 9.14). SAM/BAM-files were converted in addition into BED-files, which allowed splicing and 3' end analysis of the transcripts (bamToBed in betools suite). Data analysis was done with tools and software mentioned above (Section 9.16.3).

¹⁵ http://support.illumina.com/sequencing/sequencing_software/igenome.html, ASM294v2

9.18 SMIT DATA PROCESSING & MAPPING

SMIT data were obtained in Fastq-format, assessed for their quality (Section 9.14) and filtered for read quality (`fastq_quality_filter -Q 33 -q 20 -p 90`, Section 9.14). The 3' end read (R1) was filtered for 3' end adaptor presence and Illumina sequencing primer with `cutadapt (-g CATTGATGGTGCCTACAG -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCACGACCTCATCTCGTATGCCGTCTTCTGCTTG -n 2 -O 18 -m 23 -e 0.11 -match-read-wildcards -discard-untrimmed)`. The SMIT read (R2) was processed in a similar way, albeit with the reverse-complement sequences (`-a CTGTAGGCACCATCAATG -a AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -n 2 -m 28 -M <read length-21> -e 0.11 -match-read-wildcards`). PCR duplicates were removed with `prinseq (prinseq-lite-0.20.4)` [Schmieder and Edwards 2011] in both, the SMIT and the 3' end read. The remaining 5 nt random 3' barcode was removed from the 3' end read using the FASTX toolkit (`fastx_trimmer -Q 33 -f 6`, Section 9.14). The SMIT read sequences were split into 76 bp reads and processed reads. The processed SMIT reads were also trimmed by 5 nt (`fastx_trimmer -Q 33 -t 5`, Section 9.14).

3' end reads were mapped with `tophat2` to the *S. cerevisiae* genome (Scer3¹⁶) with the following settings: `-p 5 -N 1 -m 0 -segment-mismatches 0 -i 30 -I 1010 -g 1 -segment-length 20 -no-coverage-search -library-type fr-firststrand -min-anchor-length 8 -min-coverage-intron 30 -max-coverage-intron 1010 -min-segment-intron 30 -max-segment-intron 1010`.

The SMIT read was mapped with `bowtie2` to predefined junctions with the settings: `-L 10 -end-to-end -N 0 -k 1 -norc`. Unmapped reads were removed and the sam-output converted into bam-output. Spliced (Scer3_EEJ) and unspliced (Scer3_EIJ) Bowtie2 indices were generated from custom fasta-files. For this, the genome sequence 50 nt upstream of the 5' SS and 60 nt downstream of the 5' SS was extracted for the unspliced index. The sequences 50 nt of the first exon upstream of the 5' SS and 60 nt downstream of the 3' SS in the second exon were fused and used to build the spliced Bowtie2 indices.

9.19 SMIT DATA NORMALIZATION AND SPLICING ANALYSIS

In the ideal case, where 3' end positions (x) were uniformly distributed along the gene, co-transcriptional splicing per position (cs_x) could be determined similarly to the SPI (Section 9.15), with S being the count of spliced transcripts at position x and U being the respective count of unspliced transcripts:

$$cs_x = \frac{S_x}{S_x + U_x}$$

This would be correct, if at least one of the two following statements would be fulfilled: a) The probability to observe a read pair is independent of the insert size or b) the insert size distribution is identical for spliced and unspliced transcripts. However, the probability to obtain a read pair depends on the insert size.

16 http://support.illumina.com/sequencing/sequencing_software/igenome.html

Furthermore, splicing removes the intron of a transcript and thereby reduces the insert size of a spliced read compared to an unspliced read at the same position by the intron length l_i . Both conditions are not met and the data had to be normalized according the probability (Pr) to observe a spliced or unspliced read at position x .

$$cs_x = \frac{S_x \cdot Pr_s^{-1}}{S_x \cdot Pr_s^{-1} + U_x \cdot Pr_u^{-1}}$$

with $Pr_s = L(x - l_i)$, $Pr_u = L(x)$,
L - insert length distribution.

Assuming that the read probability depends only on the insert length one can calculate these probabilities by deducing the insert size l and knowing how insert sizes are distributed (L). The following parameters are known: The first position x_0 for each insert of a gene is defined by the SMIT primer and is constant. Thus, the 3' end defines the insert size. For unspliced inserts this corresponds to the position x in the gene. Spliced reads with the same position x are shortened by the intron length l_i . The actual insert read lengths were determined as the distance of the SMIT read start relative to the 5' SS and the distance from the 5' SS to the 3' end position mapped onto the genome. For spliced transcript the intron length was subtracted to obtain the real transcript length. Knowing that PCR steps are exponentially amplifying and biasing for short reads, it is reasonable to expect exponentially distributed insert sizes, which was the case. To deduce the exponential decay rate a linear function was fitted to the log-transformed data from intronless genes. A rate of k and a cutoff of $t=50$ were determined in that way.

$$L(x) = e^{tk} \cdot e^{-kx}, x \in [t, \text{inf}), k = 5.8 \cdot 10^{-3}$$

Co-transcriptional splicing values (cs) could be determined in that way and the position of splicing was calculated with respect to the intron end/3' SS. For visualization data points were binned in 30 nt windows and the mean and standard deviation were plotted. The saturation of co-transcriptional splicing per gene was determined as the average fraction spliced between the last three bins in the dataset, which contained at least five positions with reads. For summary statistics saturation values and the position, where 10%, 50% and 90% of the saturation value is reached, were used. The 10%, 50% and 90% saturation positions were determined by linear interpolation between adjacent data points.

Part V

APPENDIX

APPENDIX A

A.1 INFLUENCE OF THE NUMBER OF REPLICATES ON QUANTIFICATION

The degree of correlation between the 5' and 3' splice site SPI and also 5' and 3' SS ratio gives additional information on sample quality, correspondence between cell fractions and background noise. Figure 32 shows the Pearson correlation between replicates of experiments (#1 and #2) and the respective 5' SS and 3' SS ratio. 5' and 3' SS ratio cluster better within one replicate than the 5' or 3' SS ratio between the two replicates from [Khodor et al. 2012] (Figure 32A, upper panel). The opposite is true for mRNA (Figure 32A, lower panel). Furthermore, lower correlation between nascent RNA 5' and 3' SS ratios and replicates compared to mRNA (Figure 32A-B) reflect a high degree of variation in nascent RNA samples. This underlines the importance in including three and more replicates to an experiment focusing on co-transcriptional splicing to distinguish experimental from biological variability in co-transcriptional splicing values.

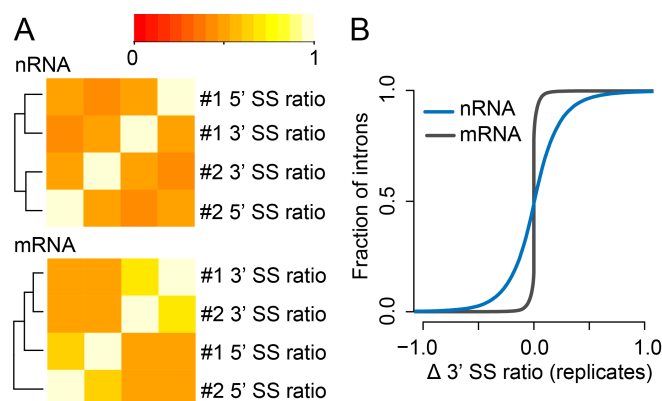


FIGURE 32 Correlation of nascent RNA-Seq splicing values between replicates for the two splicing ratios around the 3' SS and 5' SS. Nascent RNA-Seq and total mRNA-Seq data from mouse liver [Khodor et al. 2012] have been reanalyzed as described in Chapter 9. A: Heatmap of Pearson correlation coefficients between two replicates of a nRNA and mRNA sequencing experiment are shown. 3' SS ratios and 5' SS ratios for 70,273 (nRNA) and 67,203 (mRNA) constitutive, non-redundant introns are calculated as in [Khodor et al. 2011], with a lower cutoff of 200 reads in 50bp window. B: Cumulative distribution of 3' SS ratio differences between replicates for nRNA and mRNA. #1 - Replicate 1, #2 - Replicate 2.

A.2 LOW DATA CUTOFF TO REMOVE BACKGROUND NOISE

Splicing quantification relies on counting sequencing reads or calculating the sequence coverage in a pre-defined window. Depending on gene expression and

sequencing depth, many regions in the genome show very low coverage. Very low read counts are associated with high uncertainties in the splicing measurement. A minimum read count or sequence coverage is therefore usually applied. Depending on window size, read length and mode of analysis (junction reads or coverage) the cutoff has to be defined individually. To estimate the optimal cutoff, I assessed to what extent descriptive data parameters (e.g. number of introns/ exons quantified, median/ mean, 2nd/ 3rd quartile, correlation between 5' SS and 3' SS, correlation between replicates) change with increasing cutoffs. I changed the minimal read cutoff from 0 to 5,120 reads or window coverage counts. The data for the 3' SS ratio splicing measure is shown in Figure 33. A minimal sequence coverage of 200 for the 2 x 25 bp window to calculate the 3' SS ratio was chosen as optimal after considering the number of introns included, the correlation between 5' SS ratio and 3' SS ratio, the change of median and mean and the change in data distribution assessed by 2nd and 3rd quartile boundaries. Fewer introns are included in the analysis with a higher cutoff, but data correlation improves due to reduced background noise. Only minor changes in average splicing values are seen with different cutoffs.

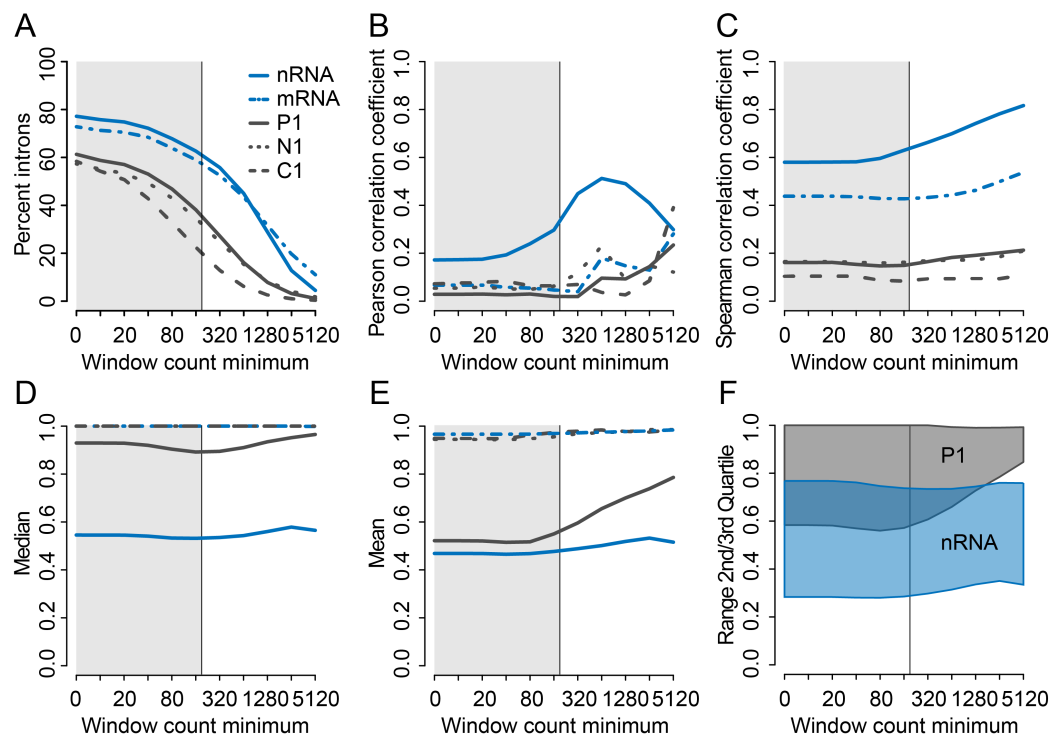


FIGURE 33 Dependence of descriptive data parameter on different minimal read cutoffs. Minimal cutoffs of total coverage in the 2 x 25 bp window around 3' splice sites ranging from 0 to 5,120 counts/ 50 bp are compared to A: the percent fraction of introns included in the 3' SS ratio calculation, B: the Pearson correlation between 5' SS and 5' SS ratio, C: the Spearman correlation between 5' SS and 3' SS ratio, D: the median, E: the mean and F: the data range between the 25% and 75% quartile for nascent RNA from mouse liver (nRNA) and chromatin-associated RNA from mouse macrophages (P1). Panels A-E represent data for all analyzed fractions (nascent RNA - nRNA, total mRNA - mRNA, P1 - chromatin-ass. RNA, N1 - nucleoplasmic mRNA, C1 - cytoplasmic mRNA)

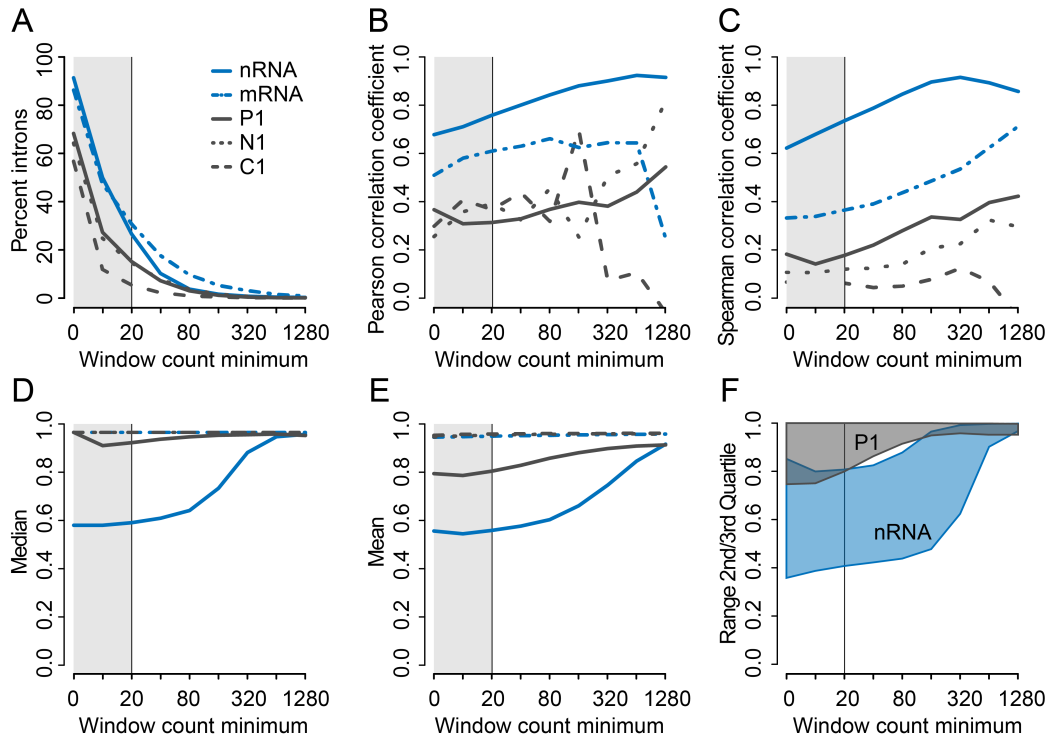


FIGURE 34 Dependence of descriptive data parameter on different minimal read cutoff for the intron-centric splicing score SPI. Minimal junction read counts ranging from 0 to 5,120 are compared to A: the percent fraction of introns included splicing calculation, B: the Pearson correlation between 5' SS and 3' SS spliced fraction, C: the Spearman correlation between 5' SS and 3' SS spliced fraction, D: the median, E: the mean and F: the data range between the 25% and 75% quartile for nascent RNA from mouse liver (nRNA) and chromatin-associated RNA from mouse macrophages (P1). Panels A-E represent data for all analyzed fractions. [nascent RNA - nRNA, total mRNA - mRNA, P1 - chromatin-ass. RNA, N1 - nucleoplasmic mRNA, C1 - cytoplasmic mRNA), SPI - splicing per intron]

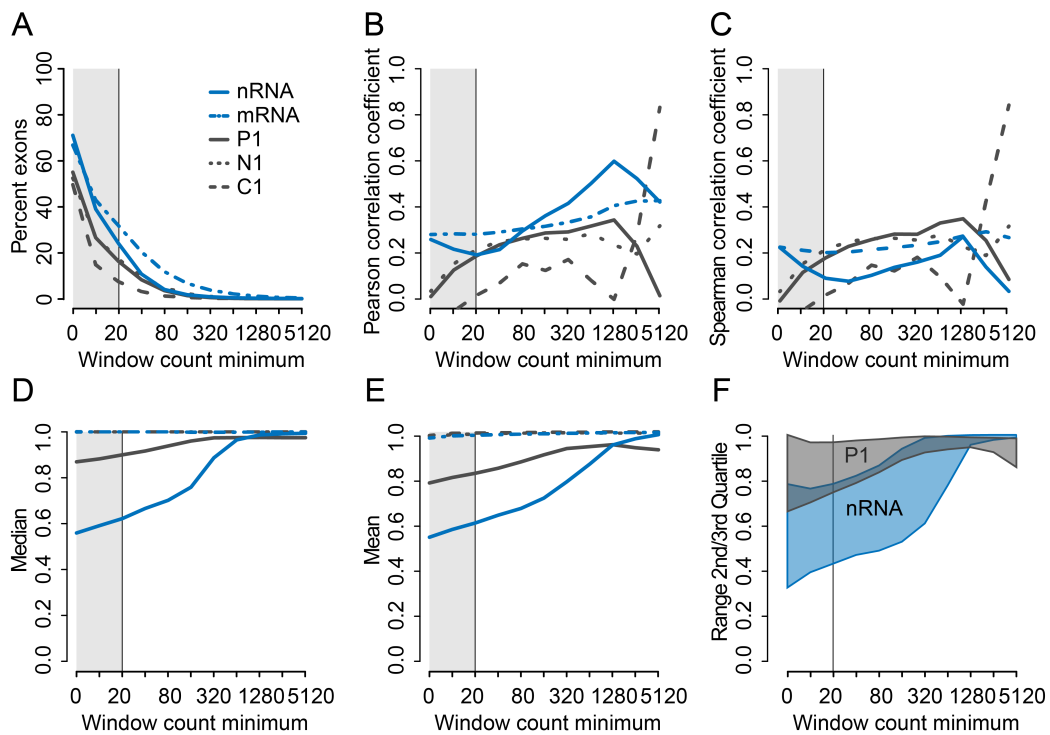


FIGURE 35 Dependence of descriptive data parameter on different minimal read cutoffs for the completed splicing index (coSI). Minimal junction read counts ranging from 0 to 5,120 are compared to A: the percent fraction of introns included in the coSI calculation, B: the Pearson correlation between 5' SS and 3' SS SPI, C: the Spearman correlation between 5' SS and 3' SS SPI, D: the median, E: the mean and F: the data range between the 25% and 75% quartile for nascent RNA from mouse liver (nRNA) and chromatin-associated RNA from mouse macrophages (P1). Panels A-E represent data for all analyzed fractions. [nascent RNA - nRNA, total mRNA - mRNA, P1 - chromatin-ass. RNA, N1 - nucleoplasmic mRNA, C1 - cytoplasmic mRNA]

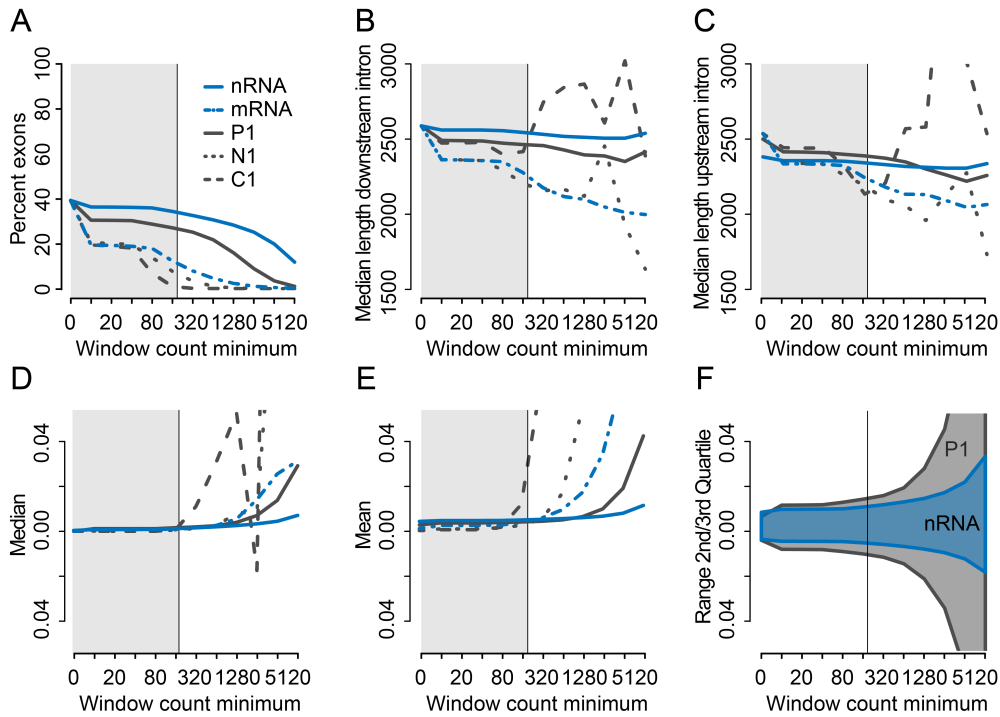


FIGURE 36 Dependence of descriptive data parameter on different minimal read cutoffs for the Intron Difference. Minimal cutoffs of total coverage in the 2×500 bp window upstream of an exon and downstream of an exon ranging from 0 to 10,240 counts/1000 bp are compared to A: the percent fraction of introns included in the intron difference calculation, B: the median downstream intron length, C: the median upstream intron length, D: the distribution median, E: the distribution mean and F: the data range between the 25% and 75% quartile for nascent RNA from mouse liver (nRNA) and chromatin-associated RNA from mouse macrophages (P1). Panels A-E represent data for all analyzed fractions. [nascent RNA - nRNA, total mRNA - mRNA, P1 - chromatin-ass. RNA, N1 - nucleoplasmic mRNA, C1 - cytoplasmic mRNA]

A.3 CORRELATION OF INTRON-CENTRIC APPROACHES BETWEEN ALL ANALYZED SAMPLES

5' and 3' splice scores were calculated for the SPI and SS ratio for all samples of the two recent publications [Bhatt et al. 2012, Khodor et al. 2012] and compared using Pearson correlation coefficients. Figure 37 shows euclidean clustering and visualization as heatmap to compare quantification approach and samples between each other.

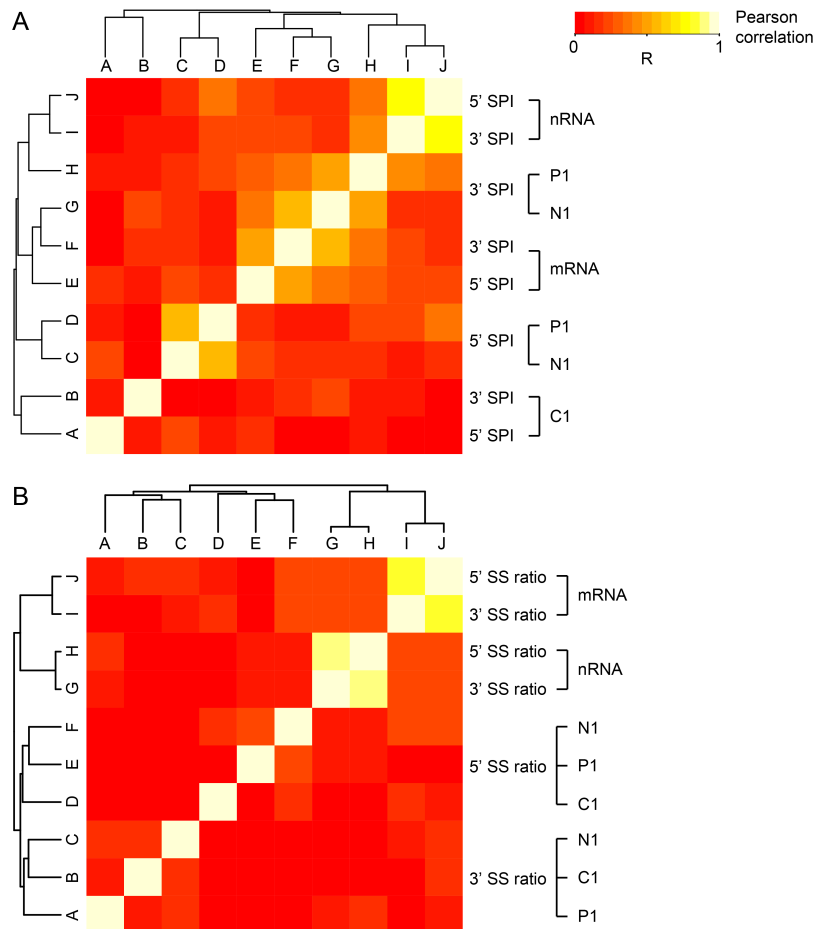


FIGURE 37 Pearson correlation of intron-centric 5' and 3' splice scores (SPI and SS ratio) and between different mouse samples. 5' and 3' splice scores correlate well for mouse liver samples; nucleoplasmic and chromatin-associated 3' splice scores correlate better than the same sample 5' splice score indicating a high fraction of polyadenylated RNA present in the chromatin fraction. Very low correlations and numbers of quantified introns are observed for mouse macrophage samples. A: Correlation for 5' and 3' splicing per intron (SPI) with $n(\text{introns}) = 1,485$. B: Correlation for 5' and 3' SS ratio with $n(\text{introns}) = 22,809$. [nRNA- nascent RNA, mRNA - total mRNA, P1 - chromatin-ass. RNA, N1 - nucleoplasmic mRNA, C1 - cytoplasmic mRNA]

A.4 HANDLING ALTERNATIVE SPLICING

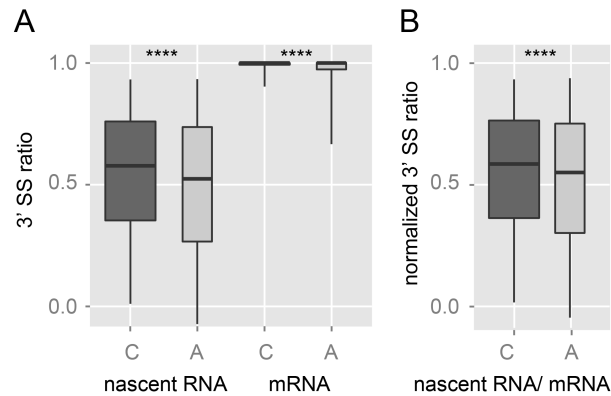


FIGURE 38 Lower pre-mRNA splicing of alternative introns. A: Boxplot comparing 3' splice site distribution for constitutive (C) and alternative (A) introns for nascent RNA and mRNA. B: Boxplot comparing normalized 3' splice site distribution (3' SS ratio [nRNA]/ [mRNA]) for constitutive (C) and alternative (A) introns. Boxwidth is proportional to the square root of the number of observations $n(\text{constitutive}) = 78,478$, $n(\text{alternative}) = 26,973$; Whiskers correspond to 95% and 5% quantiles. Asterisks indicate p-values < 0.0001 in both, two-sample Kolmogorov-Smirnov test and Wilcoxon rank sum test.

In higher eukaryotes often not only one isoform per gene is produced, but multiple isoforms. These can differ in their transcript start and end and also in their combination of exons. Fortunately, most exons of alternative transcripts are constitutively spliced and present in every isoform. Including all constitutive exons and introns in the analysis of co-transcriptional splicing might skew the distribution due to double counting. Hence, the pool of annotated exons and introns needs to be reduced to a set of non-redundant exons and introns, defined by common feature start and end positions. In order to understand what determines co-transcriptional splicing, correlations between gene architecture features (e. g. intron length, exon length, distance to gene start and end and splice site conservation), chromatin modifications, Pol II elongation rate, transcription and splicing factor association have been investigated [Carrillo Oesterreich et al. 2011]. Alternatively spliced introns and exons have been identified as less co-transcriptionally spliced than the population of constitutive introns. Our analysis confirms this observation (Figure 38A [mouse liver], $p = 3e-10$ [mouse macrophages]), using the 3' SS ratio. Significant differences have also been detected with the SPI and coSI ($p < 2.2e-16$ SPI and coSI [mouse liver], $p = 0.03$ SPI and $p = 0.003$ coSI [mouse macrophages]). The same trend of reduced average splicing for alternative introns/exons is observed for mRNA data indicating that introns and exons annotated as alternative show both reduced co-transcriptional splicing as well as post-transcriptional splicing. In order to evaluate strong co-transcriptional or post-transcriptional aspects in RNA splicing a comparison between nascent RNA and mRNA splicing patterns is required. This can be achieved by normalization to mRNA splicing values. Here, we normalize by division of nascent RNA splicing values to mRNA splicing values and thereby

correct for low nascent RNA splicing estimates due to intron retention and keep low nascent RNA splicing estimates derived from less efficient co-transcriptional splicing. The normalization reduces the difference in median co-transcriptional splicing between the alternatively and constitutively spliced intron groups as expected ([Figure 38B](#)), but does not change the overall trend.

APPENDIX B

B.1 MAPPING & CORRELATION OF *S. pombe* RNA-SEQ DATA

Nascent and mRNA single-end 76 nt Illumina sequencing data were generated in three replicates each and mapped to the *S. pombe* genome using Tophat2 (version 2.0.12) and the Ensembl *S. pombe* EF2 genome annotation. [Table 10](#) shows details on number of reads per sample and mapping efficiency.

[Figure 39A](#) shows a heatmap generated through euclidean clustering of Pearson correlation values from expression values (FPKM) estimated using Cufflinks (version 2.2.1) and the Ensembl *S. pombe* EF2 genome annotation. High correlations indicate high similarity between samples and replicates. This is true for all mRNA-Seq datasets. Nascent RNA-Seq datasets are similar among each other, but less well correlated to mRNA data. This has several reasons. First of all, mRNA expression levels are not only shaped through gene transcription (reflected by nascent RNA expression), but also through RNA surveillance and degradation (not included here). Second of all, the nascent RNA datasets are generated from polyA-depleted, rRNA-depleted samples, which generally contain higher fractions ncRNAs, which are not polyadenylated and cannot be quantified in their expression in the mRNA datasets.

[Figure 39B](#) shows the corresponding correlation to [Figure 39A](#) for intron splicing values (SPI). Clustering only differs among replicates and treatment with or without caffeine, is overall similar to [Figure 39A](#), only slightly lower. [Chapter 4](#) provides in depth analysis and a presentation of results.

EXPERIMENT	REPLICATE	READS	MAPPED	% MAPPING
nascent RNA-Seq (15' water)	1	12,060,839	11,903,670	98.7
nascent RNA-Seq (15' water)	2	11,089,042	10,982,725	99.0
nascent RNA-Seq (15' water)	3	13,611,430	13,472,446	99.0
nascent RNA-Seq (15' 10 mM Caffeine)	1	11,137,302	10,980,643	98.6
nascent RNA-Seq (15' 10 mM Caffeine)	2	13,086,301	12,943,385	98.9
nascent RNA-Seq (15' 10 mM Caffeine)	3	12,216,473	12,077,231	98.9
cytoplasmic mRNA-Seq (15' water)	1	19,358,088	19,191,208	99.1
cytoplasmic mRNA-Seq (15' water)	2	13,420,322	13,274,845	98.9
cytoplasmic mRNA-Seq (15' water)	3	16,184,316	16,017,317	99.0
cytoplasmic mRNA-Seq (15' 10 mM Caffeine)	1	14,944,348	14,803,193	99.1
cytoplasmic mRNA-Seq (15' 10 mM Caffeine)	2	15,788,091	15,629,062	99.0
cytoplasmic mRNA-Seq (15' 10 mM Caffeine)	3	14,360,517	14,232,890	99.1
nuclear mRNA-Seq (15' water)	1	23,378,994	22,524,932	96.3
nuclear mRNA-Seq (15' water)	2	23,162,556	22,162,744	95.7

TABLE 10 RNA-Seq mapping

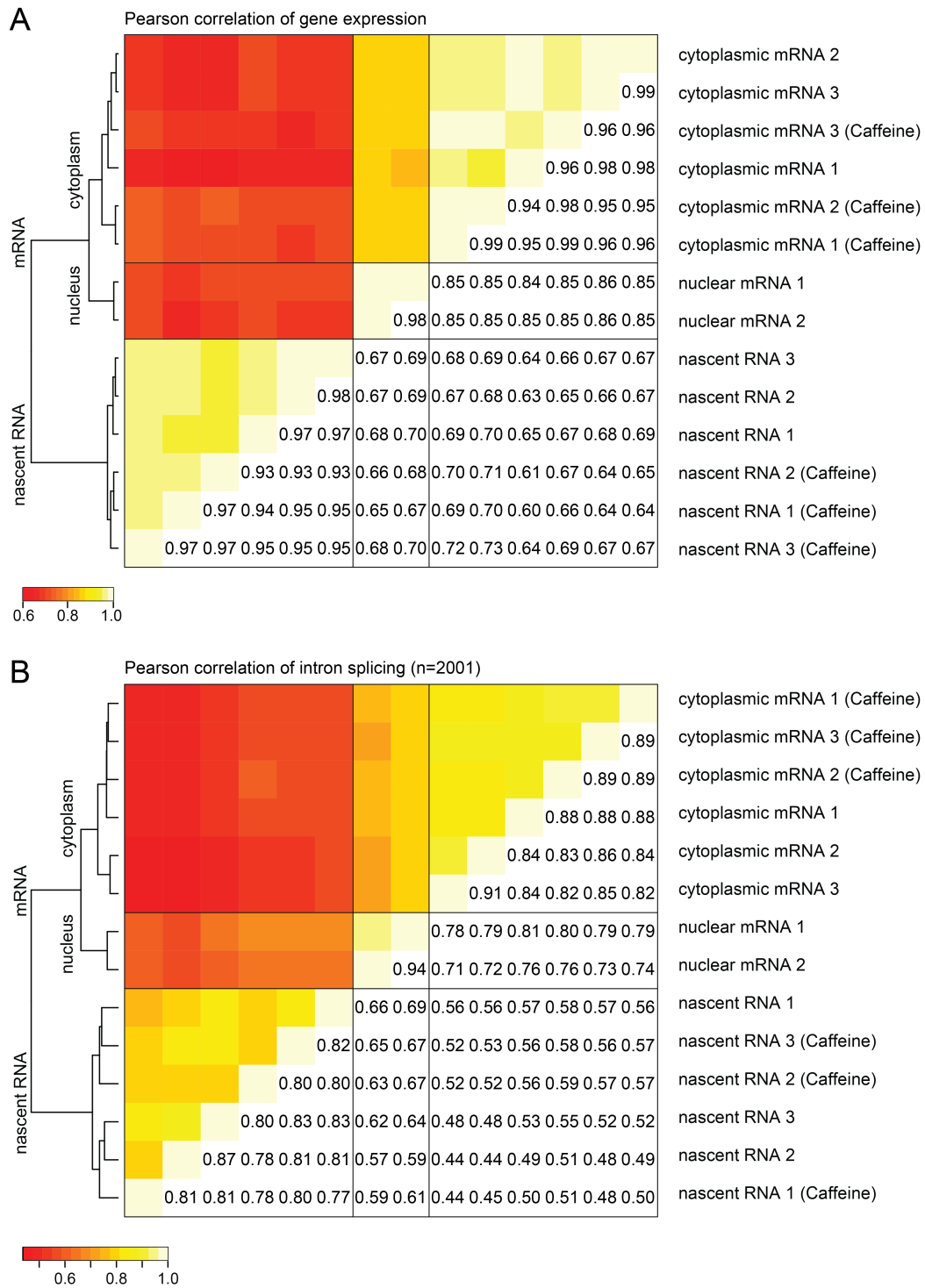


FIGURE 39 Gene expression and pre-mRNA splicing correlation between *S. pombe* RNA-Seq samples and replicates. A: Pearson correlation of gene expression values (calculated with Cufflinks) for all replicates included in the analysis of pre-mRNA splicing and splicing changes upon Caffeine treatment. B: Pearson correlation of the same samples as in A for pre-mRNA splicing. 2,001 intron were present in all datasets and used for correlation.

B.2 IMPACT OF CAFFEINE ON *S. pombe* GENE EXPRESSION AND SPLICING

I observed a significant correlation between co-transcriptional gene splicing values and mRNA expression in *S. pombe*. In order to test, if this correlation holds true, when gene expression levels are changed, I treated *S. pombe* cells with the drug Caffeine, which up- and downregulates gene expression of several hundred genes [Rallis et al. 2013]. Figure 40 shows initial growth and cell survival tests to determine the optimal drug concentration and treatment time. 10 mM Caffeine treatment for 15 min were chosen for the deep sequencing experiment.

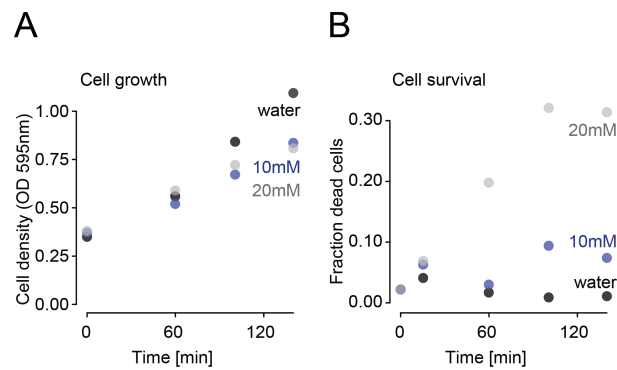


FIGURE 40 *S. pombe* cell growth and survival upon Caffeine treatment. A: Exponentially growing *S. pombe* cells in YES were treated with water, 10 mM or 20 mM Caffeine and change in cell proliferation was determined through measuring the OD at 595 nm. B: Cell death upon Caffeine treatment (10 mM or 20 mM) compared to water determined by counting cells stained with Hoechst (dead) and all visible cells in brightfield (dead + alive). The fraction of dead cells is plotted relative to the treatment time.

Three replicates of nascent RNA treated and untreated and three replicates of cytoplasmic mRNA, treated and untreated, were sequenced and global intron and gene splicing levels were calculated. Gene expression values for all samples and replicates were also determined and thus could be compared. Figure 14 in the main results section shows a modest correlation between mRNA expression and nascent RNA splicing for genes, which showed significant expression changes upon caffeine treatment ($R=0.32$). Here, I include results from correlations for all genes (Figure 41A) and the Pearson correlations between gene splicing in nascent RNA and nascent RNA expression, mRNA splicing and mRNA expression and mRNA splicing and nascent RNA expression (Figure 41B-D). Correlations are lower or not present. The very low detected correlation between nascent RNA splicing and nascent RNA expression changes, suggests that the unprocessed transcripts might decrease in stability with changes towards higher expression and thus higher nascent RNA splicing levels were detected. This remains to be tested further.

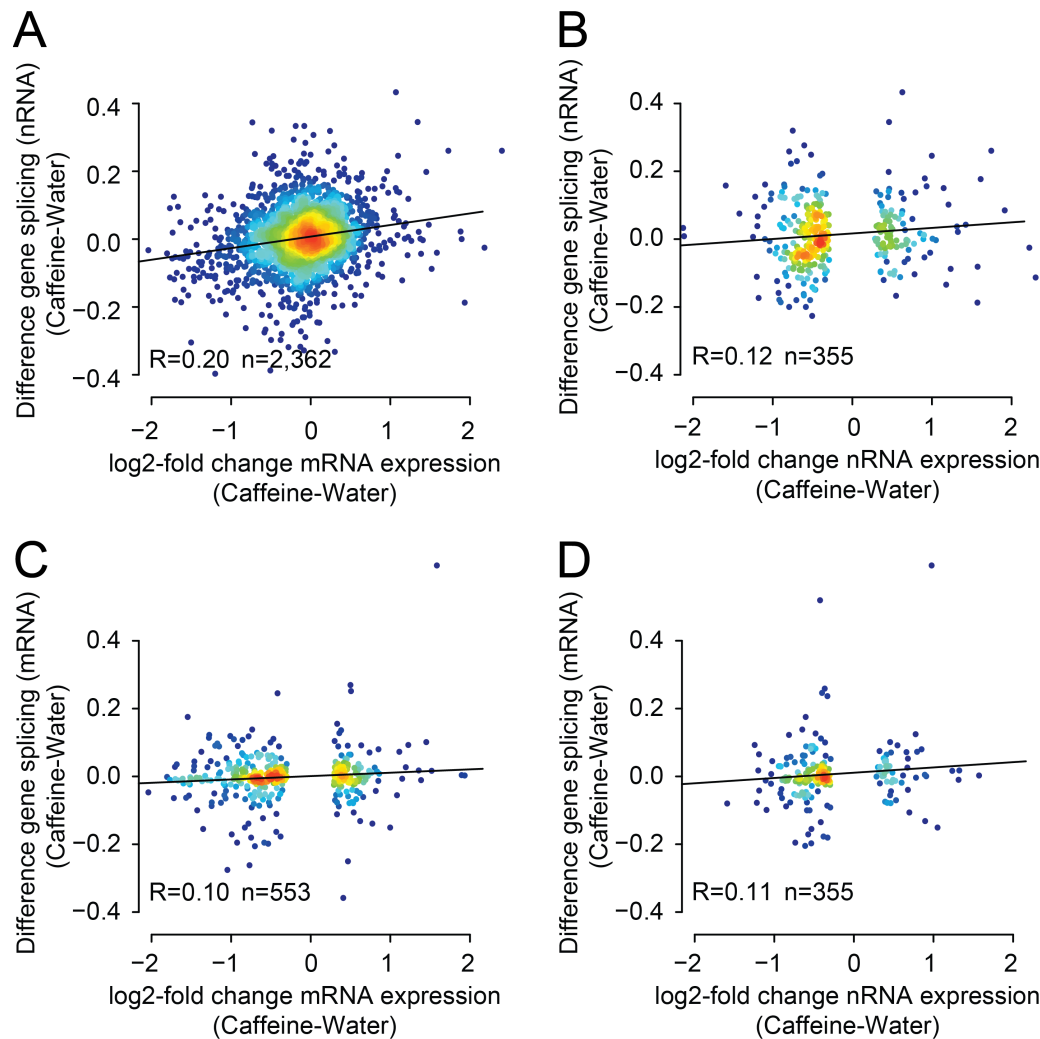


FIGURE 41 Splicing and gene expression correlation. A: Gene splicing from nascent RNA is correlated with mRNA expression changes upon Caffeine treatment. All intron-containing genes are plotted (significant and non-significant). Weak positive Pearson correlation between changes in nascent gene splicing and mRNA expression. B: Gene splicing from nascent RNA is correlated with nRNA expression changes upon Caffeine treatment. No correlation indicates no splicing change upon transcription up- or down-regulation. C: Gene splicing from mRNA is correlated with mRNA expression changes upon Caffeine treatment. No correlation is observed. D: Gene splicing from mRNA is correlated with nRNA expression changes upon Caffeine treatment. Also no correlation is observed.

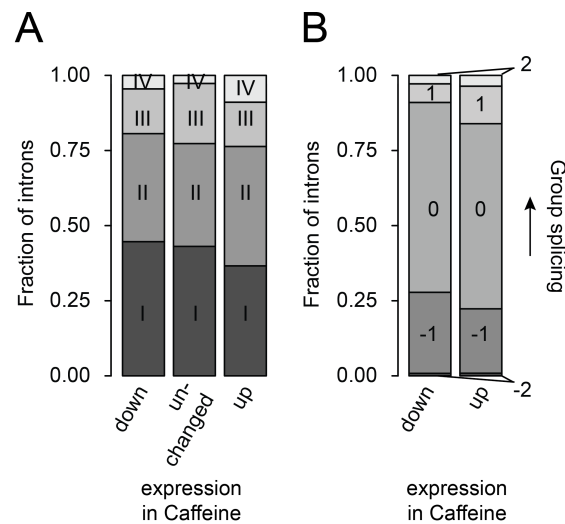


FIGURE 42 Intron groups and gene expression changes. A: Introns of genes with significant mRNA expression changes upon Caffeine treatment are compared to introns of genes with unchanged gene expression. The fractions of introns belonging to the 4 different splicing groups defined in Section 3.2.3 are shown. Introns from downregulated genes tend to be part of group I with high co-transcriptional splicing and group III & IV introns (low pre-mRNA splicing) are more prevalent in upregulated genes. B: SPIs of expression regulated genes upon Caffeine treatment are clustered again according to the new SPIs and grouped into the four respective groups of differential splicing. The fraction of introns switching groups after Caffeine treatment is shown for introns from up- and downregulated genes. A switch towards a group with higher pre-mRNA splicing is indicated with 1 or 2. Most introns do not change in group classification (0) and some especially with expression down regulation switch to lower spliced groups (-1 or -2).

In order to gain insight into which introns change their splicing levels upon caffeine treatment, I analyzed, if introns belonging to up- or downregulated genes belonged to one particular group of introns defined in Section 3.2.3. Compared to the group of introns in genes with unchanged expression, introns from downregulated genes tended to be very highly spliced (mainly group I & II), whereas especially group I introns were deriched in upregulated genes (Figure 42A). Similar to the scatter plot analysis (Figure 14) a trend was detected in the direction of group change. Although most introns in genes with changing expression did not change the group of splicing after caffeine treatment, more introns were classified with a group switch to lower splicing in downregulated genes, than upregulated genes. The opposite is also true (Figure 42B).

B.3 EXPRESSION AND INTRON SPLICING DIFFERENCES IN *S. cerevisiae* PARALOGS

S. cerevisiae contains ~6,000 genes. Many of those have have a paralog gene with similar function and sequence, which arose by genome duplication [Kellis et al. 2004]. Only few genes in *S. cerevisiae* are intron-containing (Section 1.4). 52 gene

pairs out 547 annotated as paralogs are both intron-containing¹. Most of the intron-containing genes belong to the group of ribosomal protein genes and are very highly expressed (Figure 27D). Evidence exists that *S. cerevisiae* is derived from an evolutionary ancestor, which had more intron-containing genes [Fink 1987]. In this special case of intron-containing gene evolution, the question remains, how introns and pre-mRNA splicing contribute to gene expression, so that those introns remain in genes or are slower eliminated.

In a previous study, transgenes with intron deletions were generated and effects on proliferation were studied [Parenteau et al. 2008]. The majority of introns could be removed with minor effects on cell growth. Focusing on ribosomal protein genes, the majority of introns were required for optimal cell fitness or growth under stress [Parenteau et al. 2011]. The authors also found, that the intron deletion within one copy of duplicated genes affected the expression of the paralog gene. Here, I asked about difference in gene architecture and sequence between those paralogs, and if gene expression differences under normal growth conditions are linked to splice site strength and thus pre-mRNA splicing.

I compared CDS and intron length and sequence of paralogs. The CDS length and sequence are highly similar between duplicated intron-containing genes, but the intron length and sequence differ strongly (Figure 43A-B). One third of paralogs showed very different expression profiles (Figure 44A). Gene pairs were split into to groups according their difference in mRNA expression (13 pairs high difference, 25 pairs with low or no difference in expression). Splice sites and branchpoint sequence were ranked according to abundance for all introns present in the analysis. One splice site sequence and branchpoint motif dominates in *S. cerevisiae* Table 1 and the different motifs are similar to each other. The fraction of non-consensus splice sites is higher in the group of low expressed genes compared to their high expressed paralog gene. This is not apparent, when the difference in expression is low (Figure 43C). The opposite analysis of grouping into genes with lower splice site rank versus higher splice site rank shows a trend for higher expression in genes with consensus splice sites (Figure 44B).

¹ SGD: <http://www.yeastgenome.org/>

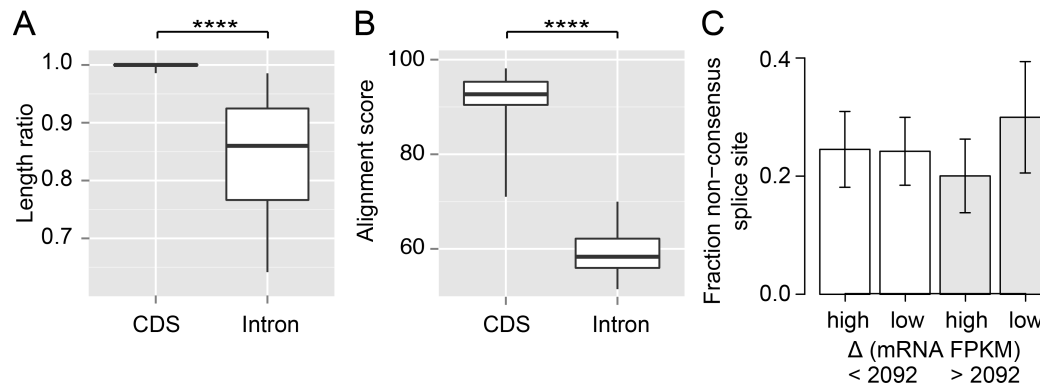


FIGURE 43 Intron-containing paralogs derived from genome duplication in *S. cerevisiae* differ strongly in intron length, intron sequence and slightly in splice site conservation possibly leading to differential expression. A: 52 gene pairs out 547 annotated as paralogs on SGD are both intron-containing. The CDS length is mostly conserved, but the intron length is not ($p < 0.0001$ Wilcoxon-rank sum test). Whiskers show 5% and 95% quantiles. B: ClustalW alignment scores (in %) of gene pairs for CDS sequence and intron sequence differ strongly, with high coding sequence conservation ($p < 0.0001$ Wilcoxon-rank sum test). Whiskers show 5% and 95% quantiles. C: 5' SS, 3' SS and branchpoint sequences of intron pairs were ranked according to their abundance. The fraction of non-consensus splice sites is significantly higher in genes with significant lower expression than their paralog. Statistical significance was assessed through bootstrapping (100 times), Student's T-test and bonferroni correction accounting for multiple testing (n.s.). WT mRNAseq data (duplicates from [Gu et al. 2015]) were remapped with Tophat2 and analyzed by Cufflinks to estimate steady-state mRNA expression values. Expression values were required to be greater than the confidence interval [Nagaraj et al. 2011] and the difference between replicates was required to be smaller than the standard deviation of the distribution of replicate differences. Generally the replicates correlate highly (0.94 Pearson correlation). Out of 52 gene pairs only 38 could be quantified in their gene expression. Paralogs were grouped according their expression difference ($n=13$ (33% of paralogs) with a difference in FPKM $> 2,092$, $n=25$ (66% of paralogs) with a difference in FPKM $< 2,092$). Figure 44A shows the distribution of expression differences between paralogs.

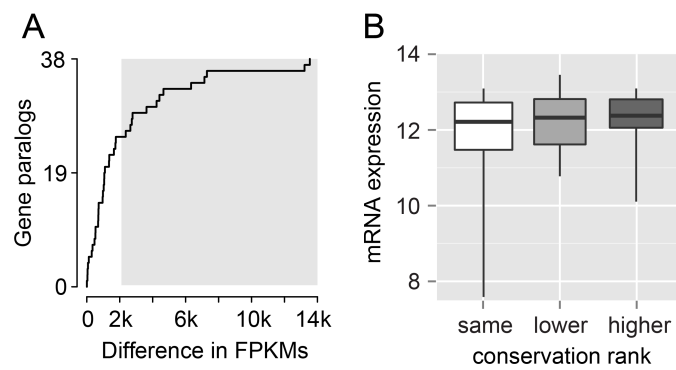


FIGURE 44 Expression differences between intron-containing paralogs in *S. cerevisiae* and splice site conservation differences. A: Cumulative distribution of mRNA expression differences between intron-containing paralogs in *S. cerevisiae*. The third of genes with highest differences (gray box) is analyzed with regard to the number of non-consensus splice sites in Figure 43C. B: Genes with higher splice site conservation than their paralog gene are slightly higher expressed (n.s. Wilcoxon-rank sum test). Boxplot reflecting mRNA expression distributions for paralogs with no difference in splice site conservation, lower and higher splice site conservation than the associated paralog. Whiskers show 5% and 95% quantiles.

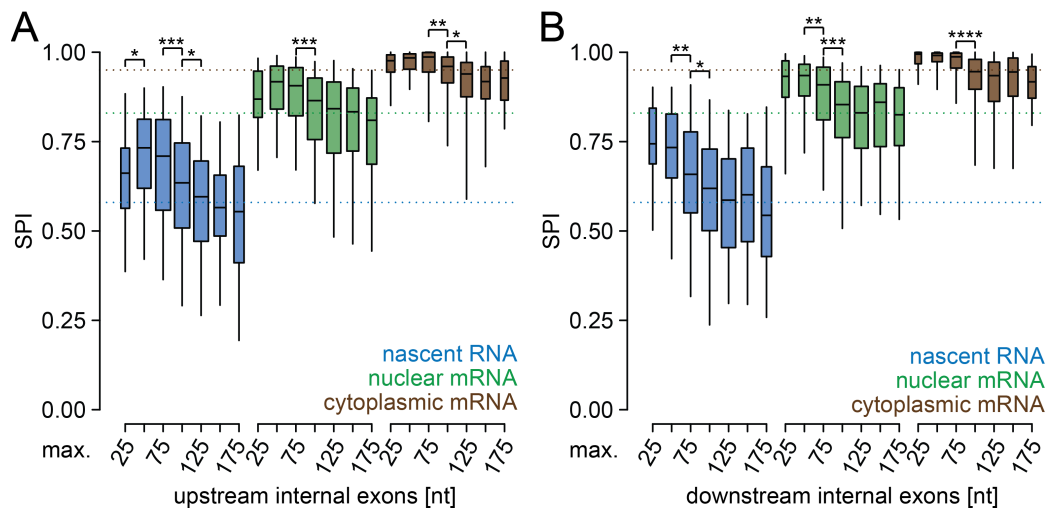


FIGURE 45 Internal intron splicing and adjacent exon length. A: Intron splicing values (SPI) are grouped according to upstream internal exon length. Internal introns with short upstream exons (<25 nt) are spliced less efficiently than introns with longer internal upstream exons. The optimal length lies between 25 and 80 nt. B: Intron splicing values (SPI) are grouped according to downstream internal exon length. Internal introns with downstream exons (<75 nt) are spliced better than introns with longer internal downstream exons.

Asterisks indicate significance of direct neighbors according the wilcoxon-rank sum test ($p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***, $p < 0.0001$ ****); Boxplot whiskers correspond to 95% and 5% quantiles. Boxwidth is proportional to the square-roots of the number of genes per group.

B.4 SPLICING OF SHORT INTERNAL INTRONS

The analysis of pre-mRNA splicing patterns and gene architecture in *S. pombe* identified especially high co-transcriptional for internal introns (Section 4.2). I detected a strong correlation of co-transcriptional with the length of internal exons. Longer internal exons tend to be spliced less efficiently (Figure 15). Here, I investigated whether there is an optimal internal exon length. I focussed on short internal exons and grouping them according to their length in intervals with 25 nt difference from 0-175 nt. Few very short internal upstream exons are significantly less well spliced co-transcriptionally than the next group of longer internal exons. This is not seen for downstream internal exons. Co-transcriptional splicing values become similar to the average *S. pombe* intron with a internal exons longer than 100 nt (Figure 45). The optimal internal exon length for co-transcriptional splicing seems to be in the range of 20-100 nt.

SAMPLE	CELLS	READS	PROCESSED	MAPPED	RRNA	FINAL
Sp_long_0213_1	2	87,614	79,155	78,968	35,865	25,922
Sp_long_0213_2	4	147,317	95,591	95,466	43,170	31,529
Sp_short_0413	1	47,018	35,339	35,328	10,577	17,773
Sp_WT_1	1	48,642	41,059	38,732	13,309	13,864
Sp_APLD_1	1	44,686	37,651	35,466	14,046	9,556
Sp_DPAD_1	1	44,984	37,612	35,394	9,998	13,375
Sp_magbead	1	45,817	39,237	38,025	12,739	12,753
Sc_1	4	131,574	104,300	95,085	NA	NA
Sc_2	2	67,966	58,492	56,147	NA	NA

TABLE 11 PacBio sequencing and mapping details

B.5 SEQUENCING FULL-LENGTH NASCENT RNA WITH PACIFIC BIOSCIENCES SEQUENCING

I addressed the questions, in which order are introns removed in transcripts with multiple introns and to which extent are intron splicing and polyA site cleavage at the end of transcription coupled, with profiling nascent RNA with PacBio sequencing. [Table 11](#) provides details on how many SMRT cells were sequenced per sample, read counts and mapping efficiency. In [Figure 46A-B](#) I present details on the library design and data processing. A single intron gene example, which overlaps with an annotated snoRNA, is shown. SnoRNA precursor transcripts are detected, but the majority arises from full-length snoRNA molecules, which are also the most abundant class of transcripts present in the dataset ([Figure 46C-D](#)). [Figure 46E-F](#) depict the final cDNA analysis by agarose gel electrophoresis before samples were submitted for hairpin adaptor ligation and sequencing.

Transcripts very similar to annotated snoRNAs are abundant in the PacBio dataset ([Figure 47B](#)), but snoRNAs in general are not more enriched in the chromatin fraction than in total nuclear lysate compared to whole cell lysate ([Figure 47A](#)). In general, PacBio sequencing is a valid technique to observe different forms of transcripts, but less quantitative than e. g. RNA-Seq mainly due to lower read counts and different representation of transcripts with different lengths. [Figure 48](#) show the correlation of the fraction of spliced transcripts of single intron genes in the PacBio datasets to the SPIs calculated from RNA-Seq. The correlation is low, albeit higher for diffusion loaded, non size selected samples ([Figure 48D](#)). This indicates that diffusion loading and PacBio library preparation of the whole pool of nascent RNAs is best to represent nascent RNA splicing.

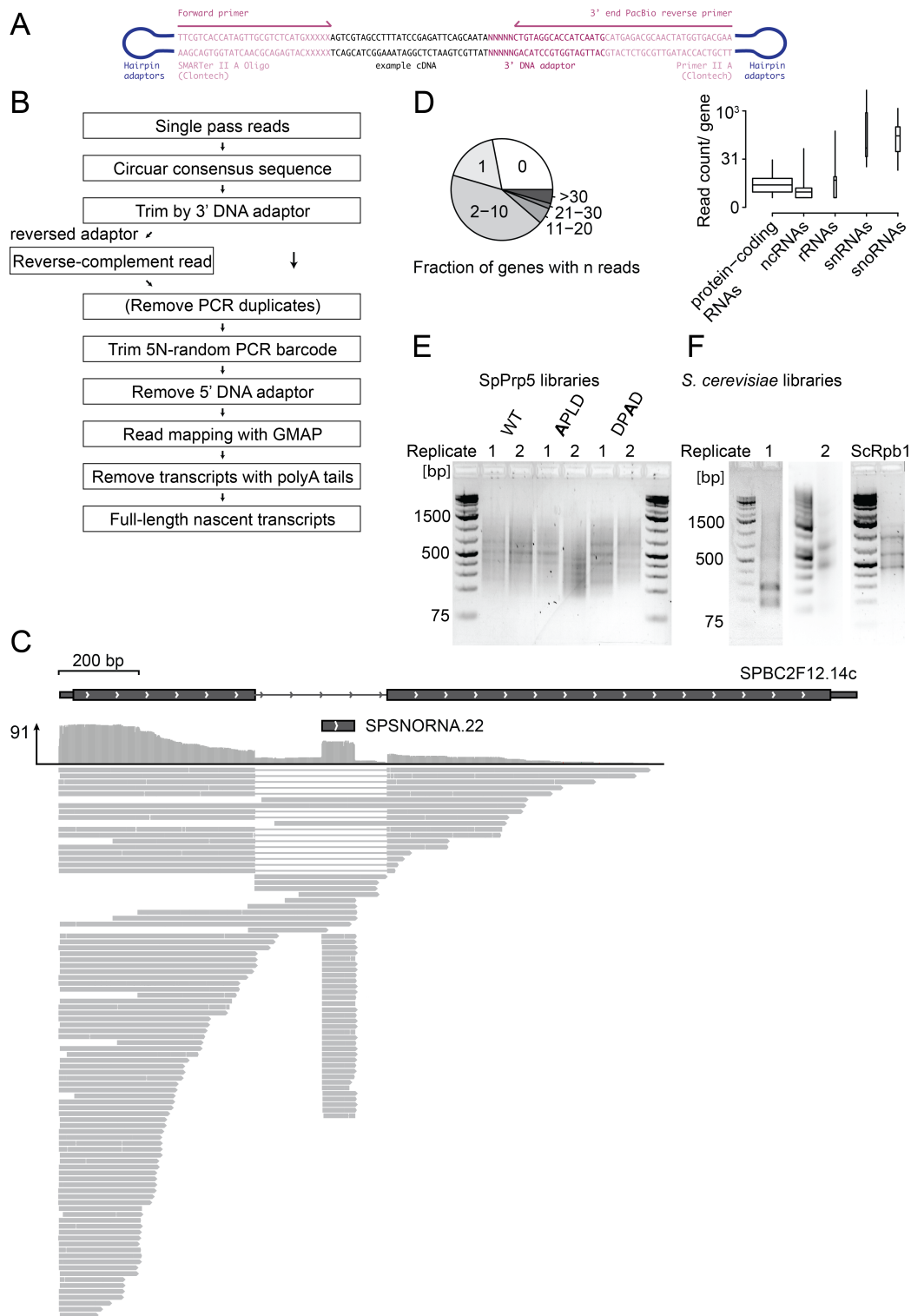


FIGURE 46 Pacific Biosciences library design and characterization (extension to [Figure 18](#)). **A:** Adaptor design and sequence details for complex, strand-specific nascent RNA Pacific Biosciences sequencing library. **B:** Detailed flow-chart of post-sequencing processing steps to remove adaptors, ensure strandedness and map transcripts back to the genome. **C:** The 124 sequenced transcripts of one intron-containing gene and its intron-encoded snoRNA are shown. **D:** 796 genes (11%) have more than 10 reads per gene with short ncRNAs, e.g. snoRNAs and snRNAs being most abundant. The left panel - pie chart of the fraction of genes with different read counts. The right panel - boxplot for PacBio read counts per gene grouping genes into different classes of transcripts. **E:** Six *S. pombe* double-stranded cDNA libraries (SpPrp5 wild-type and mutants) are shown (1.5% Agarose gels, final double-stranded cDNA). **F:** Similar to E - *S. cerevisiae* double-stranded cDNA libraries of wild-type and different replicates with gene-specific PCRs from *S. cerevisiae* Rpb1 transcription elongation mutants.

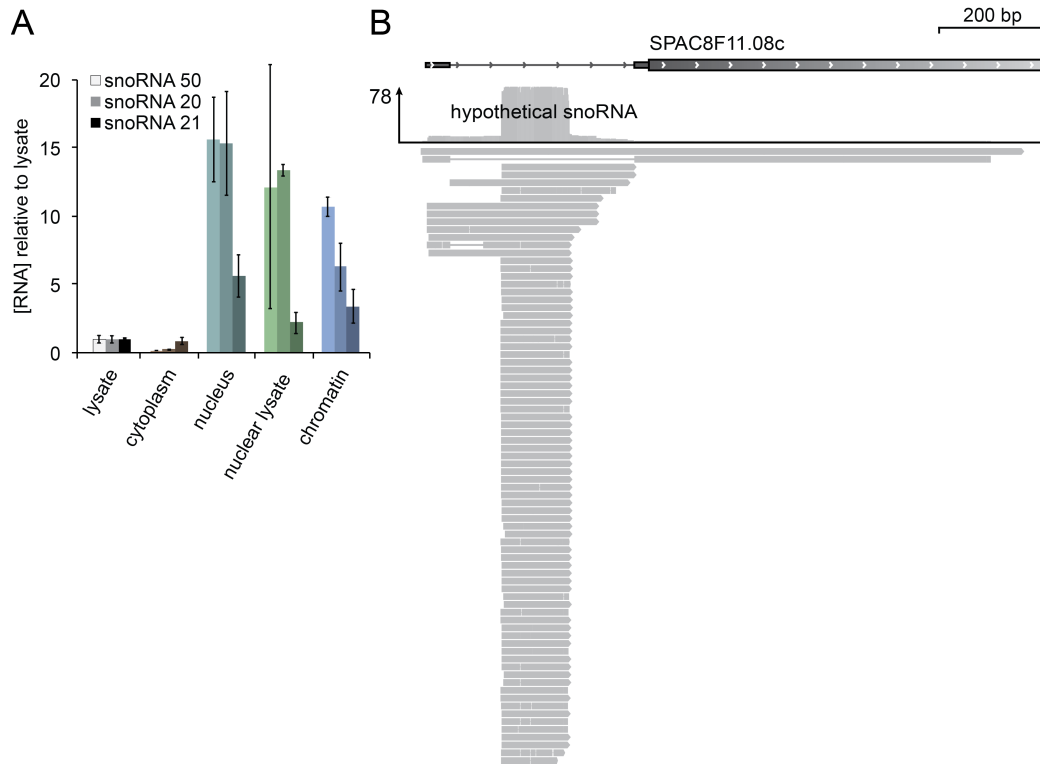


FIGURE 47 SnoRNA localization and annotation. A: RT-qPCR for 3 annotated snoRNAs in *S. pombe* in five different cellular fractions. SnoRNAs are nuclear with no significant enrichment in the chromatin pellet compared to total nuclear lysate or soluble nuclear fraction ($n=4$, SEM is shown). B: Unannotated snoRNA-like transcripts can be detected with PacBio sequencing from nascent RNA. One example is shown with short transcripts mapping to the intronic region similar to what was seen for an annotated snoRNA in Figure 46C.

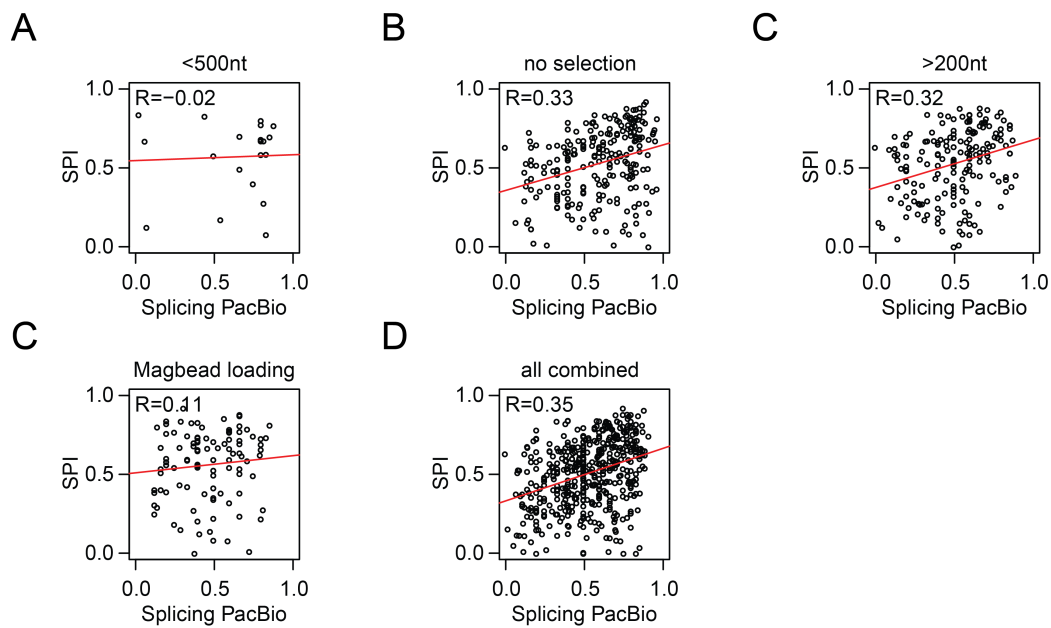


FIGURE 48 Correlation of PacBio splicing and RNA-Seq SPIs. Same splicing analysis as described in Figure 30. A-E: PacBio single intron splicing values are correlated with SPIs from nascent RNA-Seq for four different types of PacBio library (size selected RNA < 500 nt, no size selection, RNA > 200 nt and Magbead loading resulting in longer sequenced transcripts). Modest correlation to nascent RNA-Seq SPIs is observed for the combined set and the data with less size constraints.

B.6 EXAMPLES OF SEQUENTIAL AND NON-SEQUENTIAL SPLICING

About half of partially spliced transcripts in *S. pombe* are spliced non-sequentially. Most often the first intron was not removed in those transcripts detected by PacBio sequencing of nascent RNA (Section 4.3.2). This section gives examples on genes with sequential splicing and shows the gene with the most diverse splice pattern in the dataset (Figure 49A-B).

In general, first introns are less efficiently spliced compared to internal and terminal introns in *S. pombe* (Section 4.2). The group of genes, for which “in order” co-transcriptional was detected in *S. pombe*, belongs to a group of genes with very high first intron splicing. Also mRNA splicing levels are higher for those introns than for the 2nd, 3rd or 4th intron (Figure 49C, left). One hypothesis would be that “in order” splicing detected in PacBio sequencing reflects low co-transcriptional splicing of the second intron rather than sequential splicing. That does not seem to be the case in general, because differences in co-transcriptional SPIs between second and third introns are normally distributed around 0 (Figure 49C, right). However, there is a strong difference in mRNA splicing levels between those second and third introns with lower splicing levels for the second intron, which cannot be explained by non-sequential co-transcriptional splicing, but maybe post-transcriptional splicing or differences in RNA stability depending on the intron present in the transcript.

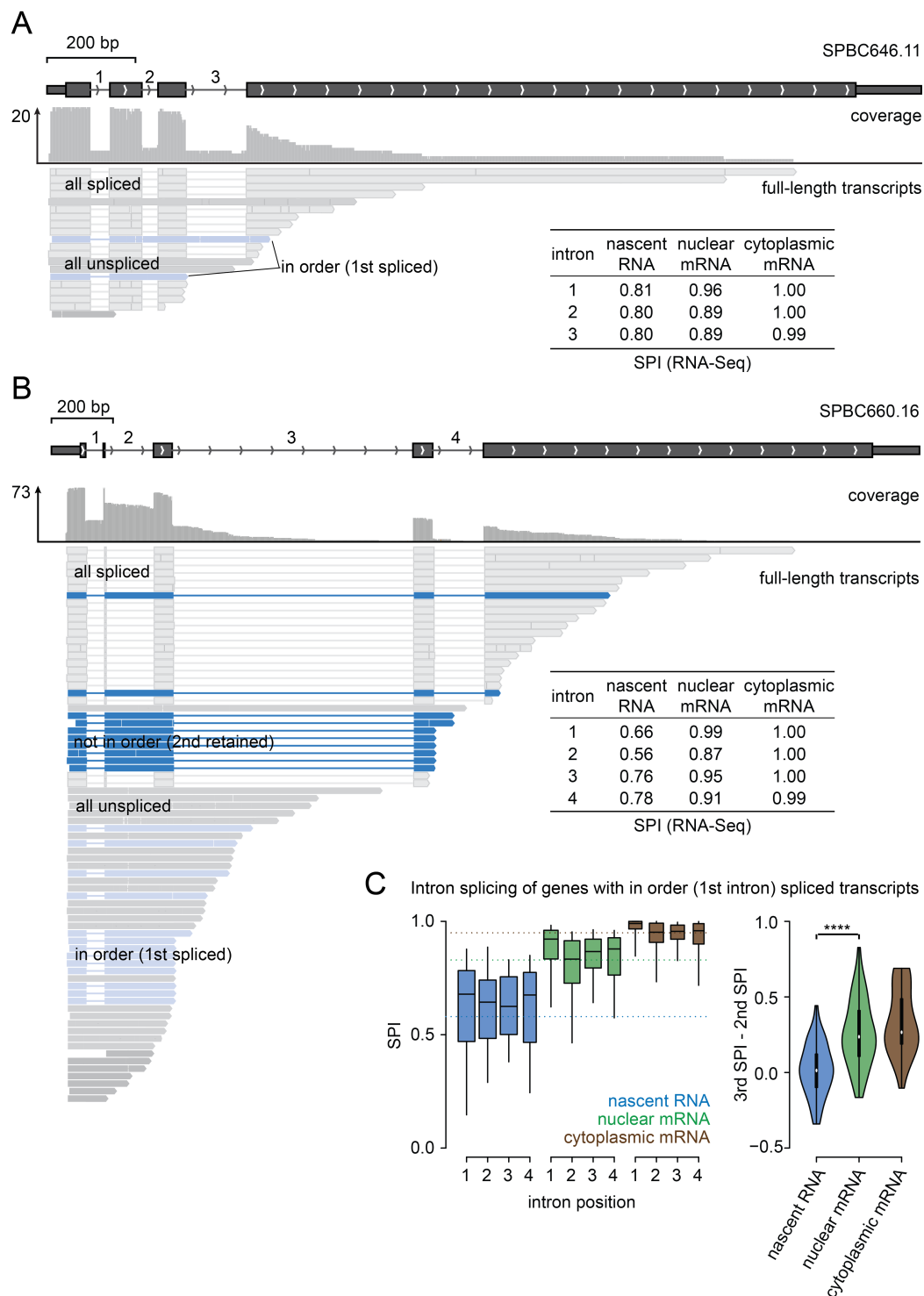


FIGURE 49 Examples of (non-) sequential splicing. Analogous to Figure 19 showing a gene example with 2 sequentially spliced transcripts (light blue) in A and the most diverse gene example in B with all 4 types of transcripts. The respective intron splicing values from RNA-Seq data (Section 3.2) are given in the table next to each example. C: Intron splicing of genes with in order spliced transcripts suggests same preference for later in order and not in order splicing. Panel 1 shows splicing value distribution for the three RNA-Seq samples of intron 1-4 of genes with more than 2 introns, which showed splicing of the 1st intron, but not the second intron in PacBio sequencing. Introns are generally better spliced than the global median (dotted lines), especially first introns. Panel 2 shows violin plots of the normally distributed splicing difference between 2nd and 3rd intron to detect, if there is a preference for in order or not in order splicing. 2nd and 3rd intron are similarly spliced in nascent RNA indicating equal chances for in and not in order splicing, but 3rd introns are significantly better spliced in nuclear and cytoplasmic mRNA (t-test $p < 0.0001$, $n=56$ (1-3), $n=36$ (4)). Boxplot-whiskers extend to the 5% and 95% quantile.

APPENDIX C

C.1 DATA PROCESSING AND MAPPING OF SMIT DATA

Single Molecule Intron Tracking (SMIT) was developed to answer the question, when co-transcriptional splicing occurs during transcription. It is a single molecule technique, which detects millions of nascent RNA molecules, their splicing end and their 3' end (Pol II position) in one experiment without disturbing the biology of the cell.

This section provides further details on the data, e. g. library preparation, processing and mapping (Figure 50, Table 12). Sample correlation is good, slightly less than in RNA-Seq replicates (Figure 51A). However, this is expected, because here I correlated counts of 3' end positions, which represent noisier data, than gene expression values, which are determined over full genes. The distribution of read counts per position is bimodal, with a many positions being detected once and a second maximum around the mean read count of 109 (Figure 51B).

DATA	Figure 50	RAW		TRIMMED	UNIQUE	MAPPED
L3229 (1)	B	2x24,403,136	R1	14,760,681	1,375,968	1,032,450
			R2	14,495,579	1,256,212	715,127
L3241 (2)	B	2x21,543,940	R1	14,473,385	1,125,643	919,075
			R2	14,160,929	1,077,313	540,874
L3242 (3)	B	2x29,883,886	R1	22,395,104	1,992,469	1,605,261
			R2	22,460,553	1,804,942	1,085,426
L3243 (4)	B	2x24,947,010	R1	19,640,064	1,612,791	1,373,847
			R2	19,582,020	1,467,551	904,492
L3244 (5)	B	2x24,869,638	R1	20,169,770	1,696,981	1,432,293
			R2	20,337,172	1,554,339	1,012,828
L3245 (6)	B	2x23,934,583	R1	19,521,725	1,813,032	1,536,728
			R2	19,500,143	1,652,540	1,056,542
WT_1	C	2x16,172,483	R1	14,111,568	841,369	540,416
			R2	12,504,185	796,793	814,595
WT_2	C	2x18,039,920	R1	16,994,804	1,056,029	692,031
			R2	15,464,227	1,005,886	1,038,635
WT_3	C	2x16,020,350	R1	13,081,726	659,859	450,989
			R2	12,133,970	633,229	603,824
FL178 (3-10)	D	2x 225,561,603	R1	166,915,586	18,362,464	12,492,935
			R2	166,175,957	17,615,913	10,232,542
SS178 (1,3-10)	E	2x 172,437,796	R1	119,912,859	6,484,316	4,189,693
			R2	90,708,286	5,189,331	4,167,924
total RNA (1)	F	2x11,626,594	R1	7,336,889	867,826	504,496
			R2	8,221,654	760,403	631,045
total RNA (3)	F	2x13,389,602	R1	7,890,083	836,399	450,038
			R2	9,543,896	741,683	644,609
total RNA (4)	F	2x11,263,821	R1	6,743,116	839,210	438,854
			R2	8,266,417	748,221	597,188
total RNA (5)	F	2x13,370,315	R1	8,468,867	910,088	533,372
			R2	9,559,725	798,819	630,654
3'PE-Seq		2x69,637,178	R1	15,087,889		14,433,990
			R2	15,087,889		12,517,606

TABLE 12 SMIT raw, processed and mapped read counts. Sample names were set by the sequencing facility or me and can refer to strain, library or replicates.

DATA	Figure 50	RAW		TRIMMED	UNIQUE	MAPPED
694 (1)	G	2x35,244,305	R1	32,353,104	3,451,037	2,760,523
			R2	29,147,111	3,110,831	2,804,517
694 (2)	G	2x13,711,567	R1	12,658,130	1,620,792	1,283,960
			R2	11,456,243	1,460,543	1,249,777
763 (10)	G	2x7,711,268	R1	6,999,998	814,210	621,495
			R2	6,622,016	746,922	673,611
763 (3)	G	2x15,069,300	R1	13,753,058	1,657,023	1,282,172
			R2	12,833,592	1,507,992	1,406,859
776 (4)	G	2x14,562,471	R1	13,562,860	1,421,626	1,156,933
			R2	12,901,383	1,317,769	1,167,253
776 (5)	G	2x7,298,982	R1	6,770,856	660,964	1,156,933
			R2	6,397,518	654,640	1,167,253
759 (6)	G	2x17,235,301	R1	15,752,959	1,909,869	1,550,565
			R2	13,527,639	1,683,549	1,505,618
759 (7)	G	2x36,135,735	R1	33,767,708	3,555,484	3,150,969
			R2	30,007,005	3,219,571	2,913,629
692 (8)	G	2x48,075,300	R1	43,937,386	4,859,674	3,999,434
			R2	39,032,206	4,312,640	3,914,138
692 (9)	G	2x21,047,079	R1	19,250,192	2,616,552	2,191,267
			R2	17,119,491	2,327,049	2,184,774

TABLE 13 Continued from Table 12. SMIT raw, processed and mapped read counts.

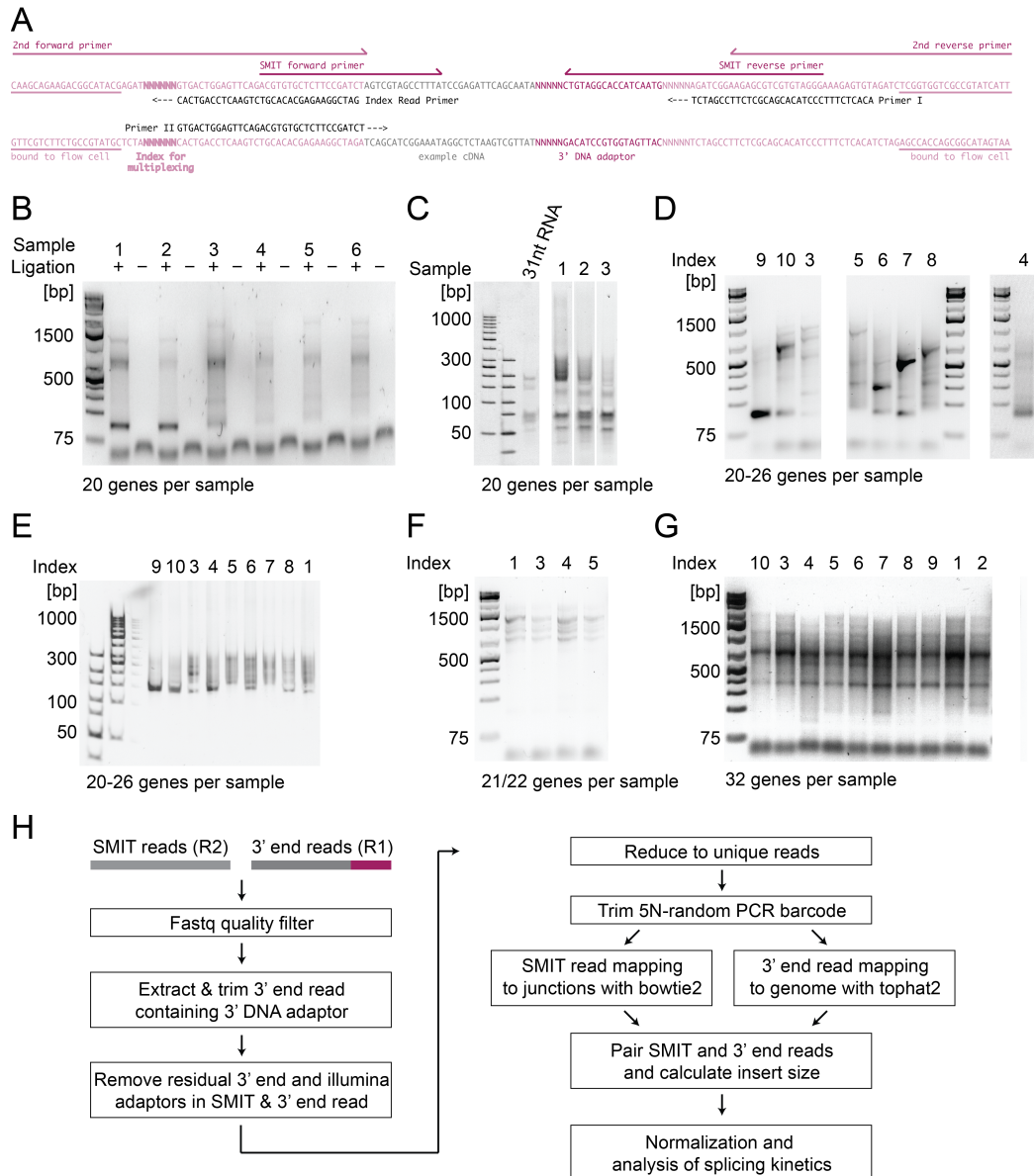


FIGURE 50 SMIT library design, preparation and data processing. A: SMIT adaptor design (extension to Figure 21) and sequence details for strand-specific nascent RNA, single-molecule SMIT sequencing library. B: Six *S. cerevisiae* wild-type SMIT libraries for 20 endogenous genes were prepared as in Figure 23. C: Three SMIT libraries on size-selected RNA (25-250 nt) for the same genes as in B. A 31 nt RNA was used for a defined control library. D: One SMIT library for 178 endogenous genes. Final PCR was done sequentially for groups of 20-26 genes grouped according to their expression levels. Each group carries a defined multiplexing index. E: Size-selected SMIT libraries for the same genes as in D and similarly prepared to C. F: Four SMIT libraries on 21-22 genes from polyA(-) total RNA. G: Ten SMIT libraries for 32 genes on different *S. cerevisiae* strains with different Rpb1 alleles. H: Detailed flow-chart of post-sequencing processing steps to remove adaptors, ensure strandedness and filter for PCR duplicates. SMIT reads are mapped to custom annotation files including assayed junctions and 3' end reads are mapped with a splicing-sensitive mapper to the genome. Libraries shown in D-G were prepared by Korinna Straube.

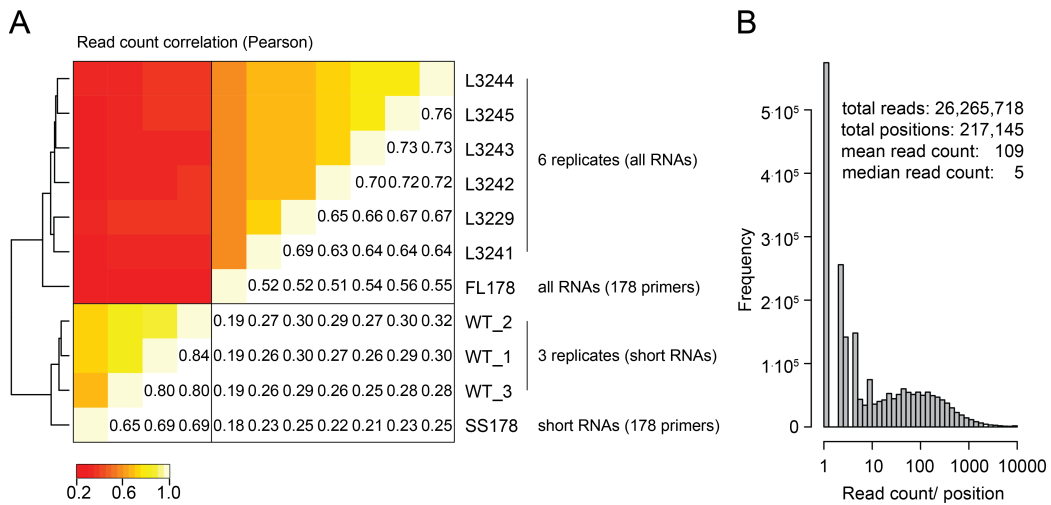


FIGURE 51 SMIT replicate correlation and read count per position distribution. A: Replicate Pearson correlation of SMIT data. The frequency of 3' ends is correlated with one another, all detected positions in any sample are included. For sample FL178 and SS178 only the subsample with Index 1 containing the same genes like the other 6 or 3 replicates is used for correlation. B: Read count distribution per position. All reads from the 11 samples in A were pooled and reads were counted per position. The highest frequency is found for one read. The distribution mean is 109 and the median 5 reads/position.

C.2 SINGLE GENE SMIT EXAMPLES

Currently, 88 genes have been analyzed with high coverage by SMIT. [Figure 52](#) shows 11 more examples. Co-transcriptional levels reach saturation for most genes within 100 nt. However, saturation levels differ substantially and profiles are also very different. Some show a stepwise increase in co-transcriptional splicing and some a more gradual increase. Also the fluctuation between splicing values per position varies on a gene by gene basis.

SMIT saturation values were correlated to co-transcriptional splicing values obtained by nascent RNA tiling array analysis [[Carrillo Oesterreich et al. 2010](#)] and 3'PE-Seq (this study, [Table 12](#)). The Pearson correlation coefficient is moderately high with highly co-transcriptionally spliced genes clustering strongly in both correlations ([Figure 53](#)). For genes with 3'PE-Seq splicing values below 0.75 lower correlation and often higher SMIT saturation values were observed. This is partly explained by the fact that those genes show later co-transcriptional splicing and thus saturation has not been reached yet in the distance window to the 3' SS, which can be assayed by 3'PE-Seq ([Section 7.3](#)).

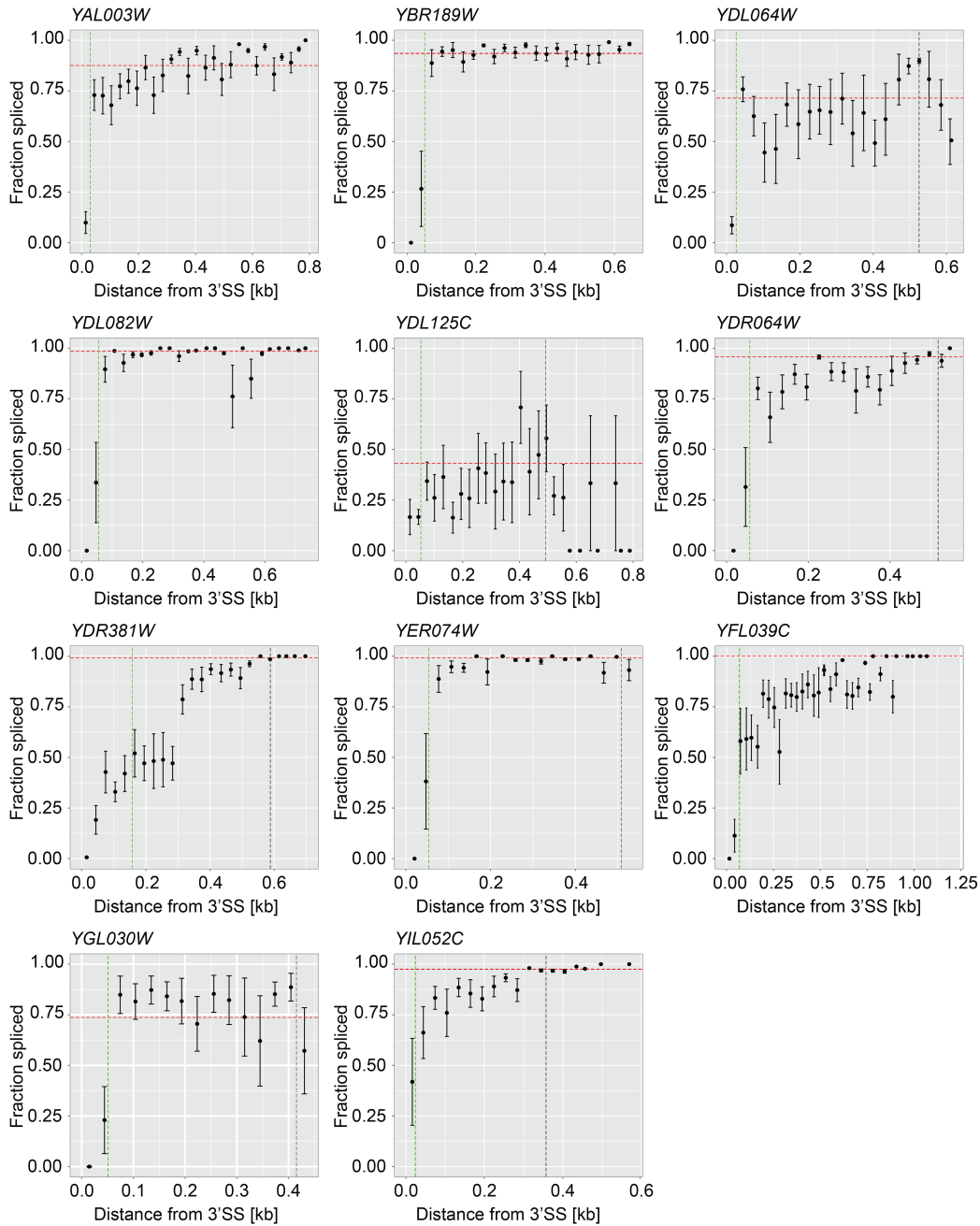


FIGURE 52 Examples of endogenous co-transcriptional splicing patterns. Eight SMIT traces showing different patterns of co-transcriptional splicing are given. Fraction of splicing is plotted with regard to distance to the 3' SS. The dashed red line indicates co-transcriptional splicing saturation. The green line marks the 50% saturation position and the dashed black line marks the terminal exon end. Data visualization in this figure and the analysis of processed and mapped SMIT data were done by Fernando Carrillo Oesterreich. Three more SMIT traces are shown in [Figure 24](#).

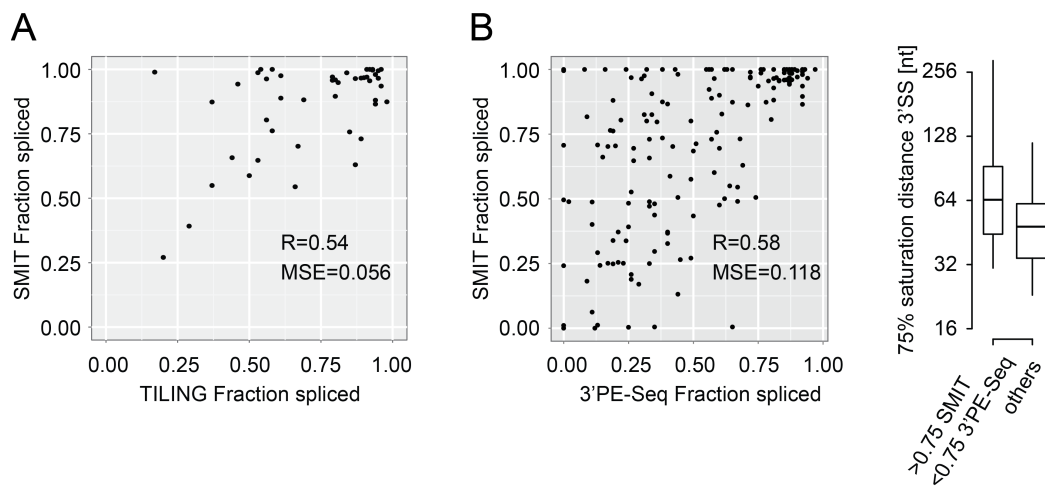


FIGURE 53 Correlation of SMIT saturation values, tiling array data and 3' PE data. A: Correlation of SMIT saturation values and tiling array data. Pearson correlation (R) and mean squared error (MSE) are given in the figure. Data visualization in this figure and the analysis of processed and mapped SMIT data were done by Fernando Carrillo Oesterreich. B: Correlation of SMIT saturation values and 3' PE-Seq data. Pearson correlation (R) and mean squared error (MSE) are given in the figure. 3' PE-Seq data have been generated from fragmented 3' end ligated RNA, thus splicing values correspond to average splicing values ~ 200 nt downstream of the 3' SS. Boxplot shows that genes not correlating well in SMIT and 3' PE-Seq have later 75% saturation values (Wilcoxon-rank sum test $p < 0.001$), thus 3' PE-Seq reflect to pre-saturation splicing values for those genes. Boxplot-whiskers extend to the 5% and 95% quantile.

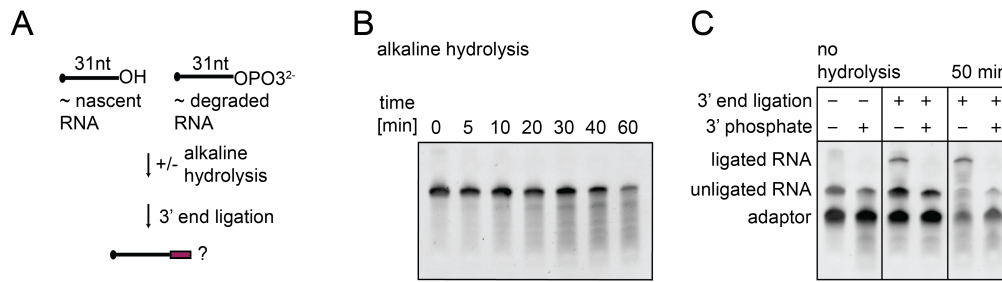


FIGURE 54 *In vitro* analysis of ligation to degraded RNA. A: Scheme of the experiment to test, if degraded and 3' phosphorylated RNA can be ligated to the 3' end DNA adaptor. B: 10% TBE-Urea PAGE of the time course of alkaline hydrolysis in Na₂CO₃/NaHCO₃, pH 9.2 fragmentation buffer [Churchman and Weissman 2011]. C: 10% TBE-Urea PAGE of (+/-) alkaline hydrolysis and 3' end ligation of 3' phosphorylated RNA and 3' hydroxylated RNA. Synthesized 3' phosphorylated RNA and 3' phosphorylated RNA hydrolysis products cannot be ligated.

C.3 SMIT IS NOT BIASED BY DEGRADED RNA

In the course of the protocol development, one worry was that SMIT might detect fragmented mRNA and co-transcriptional splicing levels might therefore be artificially high. Degraded or fragmented RNA would be 3' end phosphorylated due to the chemistry of acid-acetale bond (3' OH of ribose to phosphoric acid) and ester bond (5' OH of ribose and phosphoric acid). The ester bond is more susceptible to acid or basic hydrolysis. The SMIT 3' end adaptor is pre-activated with a 5' adenylyl pyrophosphoryl moiety, which contains the highly energetic pyrophosphate bond, which provides chemical energy for the ligation reaction. The truncated T₄ RNA ligase (missing the ATPase subunit) used in this assay should only be able to ligate the pre-activated, already 5' phosphorylated adaptor to RNA, which thus needs to be 3' hydroxylated. Therefore, hydrolysed RNA should not be ligatable. I could show this with an *in vitro* assay with presynthesized 3' hydroxylated and 3' phosphorylated RNA. Also hydrolysis products of the 3' phosphorylated RNA were not ligatable, indicating that hydrolysis, indeed, produces 3' phosphorylated RNAs. Thus, it is possible to conclude that degraded RNA should not be present in the SMIT assay. However, dephosphorylation of endogenous degradation products might occur, which then might be picked up by the SMIT assay.

APPENDIX D

D.1 PRIMERS *s. pombe*

The different sets of RT and PCR primers have been used in RT-qPCR to assess enrichment of nascent RNA over polyadenylated RNA (Figure 8E), in RT-qPCR to quantify snoRNA abundance in cytoplasm, nucleus, nucleoplasm and chromatin in comparison to total lysate (Figure 47A) and to validate 4 identified circular RNAs by RT-(q)PCR (Figure 17B).

PRIMER	SEQUENCE
anchored_oligo(dT22)	TTTTTTTTTTTTTTTTTTTTTTTTTVN
13_SPACUNK4.10_ppA	TCATCGTTGTCGTTTTACGAA
42_SPAC20G4.06c_ppA	CCGACTCATCCGTTATCACA
43_SPBC1773.10c_ppA	ACGCTAACTGACTCGCACCT
37_SPACUNK4.10_fwd	CTTCCCAATTTGGTTCCTGA
38_SPACUNK4.10_rev	TGAACGACCGTATAACATAAGCA
6_SPAC20G4.06c_fwd	CTCCTGATGTTGCTCCCATT
36_SPAC20G4.06c_rev	CGGAGAAATCAGTTGCTTGG
18_SPBC1773.10c_fwd	GCTGCTTATAAACGCGAAGG
19_SPBC1773.10c_rev	ACCAAGCGAGGAATCTTTCA
92_SPSNORNA.50_fwd	TTGTGAATCCACGTGCAACT
93_SPSNORNA.50_rev	CCCGGCTTAATTGGTGTCTA
94_SPSNORNA.20_fwd	ATGGTTGGCGTGTAGAGGTT
95_SPSNORNA.20_rev	TCAATTTTCATGGCAAGACG
96_SPSNORNA.21_fwd	TTCCATTGAACATTCGCAGT
97_SPSNORNA.21_rev	CAAAGGAAGGACTATGCACGA
145_SPBC345.06_ex2_fwd	AAGCACCTCAAGCCAAACCT
146_SPBC345.06_ex2_rev	GGATGGGATCAAAGTGCCG
145rc_SPBC345.06_ex2_fwd	AGGTTTGCTTGAGGTGCTT
146rc_SPBC345.06_ex2_rev	CGGCACTTTTGATCCCATCC
147_SPBC16G5.05c_ex2_fwd	AGGAGCTCATCTCCGTCCAT
148_SPBC16G5.05c_ex2_rev	TGGGACGAACACAGTAGTGC
147rc_SPBC16G5.05c_ex2_fwd	ATGGACGGAGATGAGCTCCT
148rc_SPBC16G5.05c_ex2_rev	GCACTACTGTGTTTCGTCCCA
139_SPAC926.09c_circ_fwd	CGGTAAAATTGGCGGCCATT
140_SPAC926.09c_circ_rev	AGGACTCGTCTAACAAAGCGG
151_SPAC926.09c_CDS_f	AGGTTTCATGCCGTCGTTTCAT
152_SPAC926.09c_CDS_r	AGCACCTGGATCAGTTCAGC
141_SPBC1815.01_circ_fwd	ACGCATAGCCTCGGAGAAAG
142_SPBC1815.01_circ_rev	GTTGGGGTGCTGAGACCTAC
153_SPBC1815.01_CDS_f	TCAACGTCTTGAACGGTGGT
154_SPBC1815.01_CDS_r	TCCGAGGAGGCAACATCAAG

TABLE 14 *S. pombe* primer sequences

D.2 PRIMERS FOR SINGLE MOLECULE INTRON TRACKING

This section includes tables of primer sequences used in the SMIT assay. The first entry in the list of first exon SMIT primers contains the general sequence necessary to add to the primer 5' end for library preparation.

PRIMER	SEQUENCE
SMIT_3'end_DNA_adaptor	/5rApp/NNNNN- CTGTAGGCACCATCAAT/3ddC/ CATTGATGGTGCCTACAG
SMIT_RT	
SMIT_1st_5N_rev	TTCCCTACACGACGCTCTTCCGATCT- NNNNNCATTGATGGTGCCTACAG
SMIT_2ndPCR_rev	AATGATACGGCGACCACCGA- GATCTACACTCTTT- CCCTACACGACGCTCTT
SMIT_2ndPCR_fwd_indexN	CAAGCAGAAGACGGCATAACGAGAT- nnnnnn GTGACTGGAGTTC- AGACGTGTGCTCTTCCGATCT
SMIT_index1	CGTGAT
SMIT_index2	ACATCG
SMIT_index3	GCCTAA
SMIT_index4	TGGTCA
SMIT_index5	CACTGT
SMIT_index6	ATTGGC
SMIT_index7	GATCTG
SMIT_index8	TCAAGT
SMIT_index9	CTGATC
SMIT_index10	AAGCTA
SMIT_controlRNA_f	GACGTGTGCTCTTCCGATCT AGTCGTAGC- CTTTATCCGAGATTC
1r_YHR010W_rev	CTTCACGTTGGGAAGGTTGT
4r_YGR029W_rev	GTCAAATTTGGGCTTCCTCA
8r_YKR095W-A_rev	ATGTGCATGAATGCCGTCTA
Universal_5'start	GACGTGTGCTCTTCCGATCT
190a_splitMS2_f_TGAAA	TGAAA GGATCTCGAGACTAGCAATAACA
190b_splitMS2_f_GTTGC	GTTGC GGATCTCGAGACTAGCAATAACA
190c_splitMS2_f_AGGGA	AGGGA GGATCTCGAGACTAGCAATAACA
190d_splitMS2_f_TGTGT	TGTGT GGATCTCGAGACTAGCAATAACA
190e_splitMS2_f_TCACT	TCACT GGATCTCGAGACTAGCAATAACA
190f_splitMS2_f_TATGA	TATGA GGATCTCGAGACTAGCAATAACA

TABLE 15 *S. cerevisiae* SMIT primer sequences (1): library & test primers

PRIMER	SEQUENCE
Universal_5'start	GACGTGTGCTCTTCCGATCT
1f_YHR010W_fwd	GTTCTTGAAAGCTGGTAAAGTTG
2f_YJR021C_fwd	ACAAGATGCTGCTACGAACG
3f_YHR101C_fwd	TGCAAACCAGACCAATGTT
4f_YGR029W_fwd	TATGACGAAGATGGCAAACC
5f_YLR048W_fwd	TGCCGCTAACACCCATTTAGG
6f_YER056C-A_fwd	GGCCCAACGTGTTACTTTCA
8f_YKR095W-A_fwd	AAACGGGAAAAGTCACTGGA
9f_YIL073C_fwd	AAACTTTGGTCGAGTTATGCG
11f_YKR004C	TCTTTTCCAAGAAGCACATACA
12f_YOL121C	ATGCCAGGTGTTTCCGTTAG
13f_YOL120C_intron	GGGTTTTAACCAACGCCAAT
14f_YNL301C	TGGTCAAACACTATACACTTTCCTAGC
15f_YDR059C	TCTTCTCCAAGCGTATTGC
16f_YNL265C	CTCCGTCAATGATTCCGTTT
17f_YML124C_intron	TTCCCAATTGGTCACCATC
18f_YAL012W_intronless	CGAACCCATTTCTTTGTCCA
19f_YBR152W_intronless	AGAGCATCCAGACCAAAAACG
22f_YBL050W_low	TGTCAGACCCTGTAGAGTTATTGAA
23f_YBL059W_low	GCCATGAAGAAAATGATAACTGC
24f_YBR078W_low	GCTATTCTAAGTGCCTCCGC
25f_YMR033W_low	GCTCCATTTAGGCAGGACAG
26f_YPR170W-B_high	GATCCTGAAGATGGACCTGC
27f_YJL001W_low	TGAAAAAGGGCGAAGTCAGT
28f_YNL112W_high	GTTGCTACTGATGTGGCCG
29f_YDR381W_high	AGGGACATTAAGCAGGATGC
30f_YDL064W_high	TGTGTCTACAGCGTCTTCAGG
31f_YPR028W_high	TCACTCTCAAATGAAACAATTCG

TABLE 16 *S. cerevisiae* SMIT primer sequences (2): first SMIT experiment

PRIMER	NUMBER	SEQUENCE
Universal_5'start		GACGTGTGCTCTTCCGATCT
YHR010W	1	GTTCTTGAAAGCTGGTAAAGTTG
YJR021C	2	ACAAGATGCTGCTACGAACG
YHR101C	3	TGCAAACCAGACCAATGTT
YGR029W	4	TATGACGAAGATGGCAAACC
YLR048W	5	TGCCGCTAACACCCATTTAGG
YER056C-A	6	GGCCCAACGTGTTACTTTCA
YHR079C-A	7	GAATGGCAACCTGATAAAGATGT
YKR095W-A	8	AAACGGGAAAAGTCACTGGA
YIL073C	9	AAACTTTGGTCGAGTTATGCG
YKR004C	11	TCTTTTCCAAGAAGCACATACA
YOL121C	12	ATGCCAGGTGTTTCCGTTAG
YNL301C	14	TGGTCAAACCTATACACTTTCCTAGC
YDR059C	15	TCTTCCTCCAAGCGTATTGC
YNL265C	16	CTCCGTCAATGATTCCGTTT
YBL050W	22	TGTCAGACCCTGTAGAGTTATTGAA
YBL059W	23	GCCATGAAGAAAATGATAACTGC
YBR078W	24	GCTATTCTAAGTGCCTCCGC
YMR033W	25	GCTCCATTTAGGCAGGACAG
YPR170W-B	26	GATCCTGAAGATGGACCTGC
YJL001W	27	TGAAAAGGGCGAAGTCAGT
YNL112W	28	GTTGCTACTGATGTGGCCG
YDR381W	29	AGGGACATTAAGCAGGATGC
YDL064W	30	TGTGTCTACAGCGTCTTCAGG
YPR028W	31	TCACTCTCAAATGAAACAATTCG
YLR054C	32	AAGAAAACCAATTAACGTGCTTA
YAL001C	33	AAAATTGCTTCAAATAAGGGAA
YAL003W	34	GCTGACAAGTCATACATTGAAGG
YAL030W	35	GGAAGTGCAGGAACTACAAGC
YBL018C	36	AATGGCAATATTTCAAGTTATCAAT
YBL026W	37	TGACCAAGAAGTGGTCGTAGAG
YBL040C	38	GCAATGAATCCGTTTAGAATCTT
YBL059C-A	39	AGTTAGAAGCTGAACGTTTTTATT

TABLE 17 *S. cerevisiae* SMIT primer sequences (3a)

PRIMER	NUMBER	SEQUENCE
YBL072C	40	CAAGCGTCATTAATTTTCTATTACA
YBL087C	41	CGGTGCTCAAGGTACTAAGTTT
YBL091C-A	42	AGAAAAGCTGGTGTTCAAGG
YBR048W	44	AATTAAGTGTCAATCTGAAAGAGC
YBR082C	45	ACGTATTGCTAAAGAATAAGTGATCT
YBR084C-A	46	GATTAAGAATACTCCAAAGCAAAAAT
YBR089C-A	47	CGCACGACAAGAGTACTAATAAT
YBR090C	48	CACCAAGCAGAAAAACGACTT
YBR119W	50	AACCTAAAGAAACCATGTCAGC
YBR181C	51	GAATAGACGACTGAGCCATCA
YBR186W	52	AAAATTGCTATCAAATGCTCG
YBR189W	53	AAGACTAAGCAACAATGCCAA
YBR215W	54	GATTGGAAATGGACCAAAAAG
YBR219C	55	GCCAATCAGTTCCTGGAGAT
YBR230C	56	CAGCATCTCATAATATGTCTGCAA
YBR255C-A	57	CATTATGACCCCAAACTGTAAAA
YCL002C	58	AGGGCTTGGCAGGTTTTT
YCL012C	59	ACACCGGAAAACCAAAGTT
YCR028C-A	60	CAATGAACCTCAAATCAATTTTT
YCR031C	61	CAAGAACCCGCCATGTCTA
YDL012C	62	TCTGCCTCCAAACAAAGC
YDL029W	63	AATGGACCCACATAATCCAA
YDL075W	64	CACCATTAAGTTGCACAAAAGA
YDL079C	65	GGTATATTAACCAAGGAAAGGAC
YDL082W	66	GCAGGAATCGTACACAATGG
YDL083C	67	CTGCCGTCCCAAGTGTCC
YDL108W	68	AAAGTGAATATGGAGTACACAAAGG
YDL115C	69	GGATTCTGTGCAGGCATTAT
YDL125C	70	CTTGATGCTGCCTGTATTTTT
YDL130W	71	CCGCTGGTGCTAATGTCG
YDL136W	72	CGATAAAGAACCAAATAGGACTAAAAA
YDL191W	75	CCGTAGAATAGGTACAGTGAGACA
YDL219W	76	CGTCGATTCAAAGTTATTTCAAG
YDR005C	77	CCTAAAGAATCACGACAATGAAA
YDR025W	78	CCGAGAGAGCTTTCCAAA
YDR064W	79	AAATGGGTCGTATGCACAGT

TABLE 18 *S. cerevisiae* SMIT primer sequences (3b)

PRIMER	NUMBER	SEQUENCE
YDR092W	80	ATCATTACCCAAGAGAATAATCAAG
YDR129C	82	CCAAAACACAATGAATATTGTCAAA
YDR139C	83	TCAACAAAGACTTATATTCCAGGG
YDR305C	84	AGTTTCTTGTAAGTGAACAAGTTTTCT
YDR318W	85	GCAGGACATTGAATCTTTACTCA
YDR367W	86	CATCTCTTCTAACCCTTCCAAA
YDR381C-A	87	TGTCAAATCCATTTCAAATATAGG
YDR397C	88	ATGTGTCGCTTCCCAAGG
YDR471W	91	AATTTTTGAAAGCAGGTAAAGTTG
YDR535C	92	AACGGCACTTATTTTCGTCC
YEL012W	93	GGAGCGTAATACGAAAGATGAG
YER003C	95	AGCTGTTCAAGTTAGATGCAG
YER007C-A	96	CAGCAGAGAGGAAATGTTTAAGAA
YER014C-A	97	TGTTTTGTACCTGGGATAGCTG
YER044C-A	98	CATCGGAAGTTGACTGGATAA
YER074W	99	AACCAACATAGATTAAGCAGAAATG
YER093C-A	101	GGGCCATAAAAGTACGAAAAT
YER117W	103	TGCTCAAGGTACAAAATTCAGA
YER133W	104	CTAGAGTTAGAAGCCCCAATTA
YER179W	105	CTGGTGGGATATACACAGTCAA
YFL031W	106	TGACAATTGGCGTAATCCAG
YFL034C-A	107	AACCACAATGGCTCCAAAC
YFL034C-B	108	CATCATGTCCTTCTTCAACTTCA
YFL039C	109	ATTTACTGAATTAACAATGGATTCTG
YFR024C-A	110	CAAGGAGTTTAAAGAGTGAGACAAA
YFR031C-A	111	AAGAAACCATTAGATCAATAAGCAA
YFR045W	113	GCACCCGACATGGCTAAC
YGL030W	114	CAATTAATCAATATACGCAGAGATG
YGL031C	115	TGTTTGAAAAACGTGGATTAATATAG
YGL033W	116	CAGGCAAAGGCTCAGAAG
YGL087C	118	CGTTTACACAAAATGTCAAAAGT
YGL103W	119	GCACAGAGGTCACGTCTCA

TABLE 19 *S. cerevisiae* SMIT primer sequences (3c)

PRIMER	NUMBER	SEQUENCE
YGL137W	120	GGACACGATGAAGTTGGATATAAA
YGL178W	121	CGATTTTTCCAGTTTCTCTTATG
YGL226C-A	124	CAAGGTTTACAATGACTTATGAACAA
YGL232W	125	AATTCCCAAACAGGAAGAAA
YGR148C	129	GGTTTTTCACAGCTTTTTAATTTAG
YGR183C	130	AACAATAGCAATACGGACTAAAATG
YGR214W	131	CACTTAGGTGCTAGAAACGTTCA
YGR225W	132	CTCAGGCACAGCAAGTCTG
YHL001W	133	TTGAAATTATCGACCAAAGAAG
YHL050C	134	AGAAATCCAAGGTGCCTTTAG
YHR001W-A	135	TTAAACTAACCTCAATGGCG
YHR012W	136	GCACATATTCCTGATAGAGCAA
YHR016C	137	GCTTGAAAAGCGAGACCAA
YHR021C	138	AAAACACACCAGATAATTAGTGCAT
YHR041C	139	CCATCATGGGAAAATCAGC
YHR076W	140	TCGAGAGGTCCCCTTTATG
YHR077C	141	AATACATTGGACAGAAATTATGGAC
YHR097C	142	TGAAACGCCCAGGAATTAT
YHR123W	143	TTCACATATCGAAAATCTAAAGTCA
YHR199C-A	144	ATCGAATGCAGCTGGAAC
YHR203C	145	GGAAAGATGGCTAGAGGACC
YIL004C	147	CTAGGCTACACAGATGAGTTCAAG
YIL018W	149	CCACAAAGTTATTGAACAATGG
YIL052C	150	TGTTACTTTCAGAAGAAGAAATCCA
YIL069C	151	GGTACCTACAGAAGATCAATAAAAATG
YIL106W	152	TTCAATTCCATGTCTTTTCTACA
YIL133C	154	TGAACCAGTTGTTGTCATTGAT
YIL148W	155	CCAAGATTCAAACATGCAAA
YJL024C	157	CACAATGATTCATGCAGTTCTAA
YJL041W	158	AGTAATAAGCTCTGATCGTTTTGAA

TABLE 20 *S. cerevisiae* SMIT primer sequences (3d)

PRIMER	NUMBER	SEQUENCE
YJL136C	160	GGAAAACGATAAGGGTCAATTA
YJL189W	162	AACACAGATAGATCAACATGGCT
YJL191W	163	CAATAACAATTAAGAATGGCTAACG
YJL205C	164	CAGTACGCTCCCTTTCTATTAGG
YJR079W	165	GTTGTTTCTTCTTGGTCATATTTTT
YJR094W-A	166	AAACAGCATAGATAATCAAACAAAAA
YJR112W-A	167	ACAATGGTACAGCTGAGAAGAAC
YJR145C	168	GCAAAGATGGCTAGAGGACC
YKL002W	169	TCCGCAAGAGAGGTTAAAAA
YKL006C-A	170	GGTATTCTCAGACCGAATCAAA
YKL006W	171	TCGAAATTATTGACCAAAGAAG
YKL081W	172	AGCTTTGGCTATCCAATTTTATT
YKL157W	174	TGTGCGAAGTGCCAAAAA
YKL190W	177	TGGATGGTCTTTTAGAAGATACAAA
YKR005C	178	CCCTTCCTCTCCAGTTGC
YKR057W	179	GGAAAACGATAAGGGCCAAT
YKR094C	180	CCAAAGGTTCAAAATGCAAA
YLR061W	183	GAAACAATGGCCCCAAAC
YLR078C	184	AATCATATTGATTTGAGGGGG
YLR093C	185	TTATATTTTTACCAAATGAAACGCT
YLR199C	186	AAATTTTCATTTATAGCGATGCTT
YLR202C	187	GCGTTTTCTACCGTTATTATAATTTTT
YLR211C	188	TGGAATGAGTACTTTAGCGGAA
YLR275W	189	CCTTTGACAGTTGATTAGAGGAG
YLR287C-A	190	CCCATACAAAACTACGCAAA
YLR329W	192	CTACCACCTGGAATGTTGAAA
YLR344W	194	TATCAGAATGGCTAAACAATCATT
YLR388W	196	ACTTCGACAGTCAACAACAATTT
YLR406C	197	CCATCAACTTGCACAAAAGA
YLR426W	198	TCTGGTAGCTCTTGATTGAGATG
YLR445W	199	AAAATGGAAAGGCTAGCAAAA

TABLE 21 *S. cerevisiae* SMIT primer sequences (3e)

PRIMER	NUMBER	SEQUENCE
YLR448W	200	CGAAATGACTGCCCAACAA
YML024W	202	CTCGAGACTAGCAATAACAAAATG
YML025C	203	CGATAAAAAGAAATTTGGTGAAA
YML034W	205	ACCAGAAATATTATGGAGGCAA
YML036W	206	ACGAAGCGCTCATAAGAAAA
YML056C	207	TTTCTCTGGCTTCCCAGTTA
YML067C	208	TGAAGACATTTGATGCGTTTC
YML073C	209	CGAAATGAGTGCCCAAAAA
YML085C	210	CAAACAAACAATGAGAGAAGTTATTAG
YML124C	211	GAGACAATGAGAGAGGTCATTAGTATT
YMR079W	213	GCCCTAAAACACAATGGTTACA
YMR116C	214	CGGTAACGACAAAATGGTTAAG
YMR133W	215	GGAAAAGTTGAAAGACGAAGAA
YMR143W	216	CAGCTGTCCCAAGTGTTCAA
YMR194C-B	217	TCCATGCCAGAAGGAGGC
YMR201C	218	CCGAACAAAAGGCCAAACTA
YMR225C	219	ATCCCTTTGGCAAGGAAG
YMR230W	220	CAA AATTCACCAATACTTGTTTCAA
YMR242C	221	GGAAAACAGTGCGGAAAAA
YMR292W	222	GGCTCACAGAGGCTCAAA
YNL004W	223	TCAAGAACGAGGTCAGAAAA
YNL012W	224	CCACTTATTAGGGAGGCCAA
YNL038W	225	GTCTACGCTTAAGCTATTATTTCCA
YNL044W	226	TGAATCAGTTAGGAGCTTTAGCC
YNL050C	227	CATTGAAATTCAGTATAAAAATGTCTG
YNL069C	229	CAACCAGTCGTTGTTATTGATG
YNL096C	230	TTGCAAATCAAATCTATCAGAGAA
YNL130C	231	GAGTAGTTTGGGAAACTTGAAGC
YNL138W-A	232	GAAACAGTGAAATTAAGAAAAGAAATG
YNL147W	233	CATGCATCAGCAACACTCC
YNL246W	234	CAAGAAAACGAGAACGAGCA
YNL302C	235	TGGCAGGTGTTTCCGTTAG
YNL312W	236	CTAGTTTAAGCATATACATAATGGCAA
YNR053C	237	TGGTACCTACCTGGGTTGC
YOL047C	238	GTCGTCGCACCAGATCAT
YOL048C	239	CAAGAGCATTATCTACCCATTCTT

TABLE 22 *S. cerevisiae* SMIT primer sequences (3f)

PRIMER	NUMBER	SEQUENCE
YOR096W	240	GCAATTCAAGTCCATCAGAGAA
YOR122C	241	CGCAAATTATGTCTTGGCA
YOR293W	244	AGATCCACCAATACTTATTCCAAG
YPL031C	245	AATACCAATGTCTTCTTCTCACA
YPL079W	247	GACACTAAACAAAAATGGGTAAATC
YPL081W	248	AAATACAAAAGTATACAACATGCCAA
YPL090C	249	GACAAAAGAGTGAAGACAGACTATACA
YPL109C	250	CCAAACTTGGACTTATTTGAAAG
YPL129W	251	ACCTCGGAGCTGACTGATATT
YPL175W	253	AAACAATGGGCTTCAATATAGC
YPL218W	255	GTTGGGATATTTTTGGTTGGT
YPL249C-A	257	AATACAAAATGGCTGTCAAGACT
YPR010C-A	258	AAAAATGAGACCAGCACAGTTAC
YPR043W	259	CAGGAAGACA ACTGAGACAAAAA
YPR063C	260	GCCCGACCTTTGTGTTTC
YPR098C	261	CTACGGCTCATTTGCTTTTT
YPR187W	263	CATGTCAGACTACGAGGAGGC

TABLE 23 *S. cerevisiae* SMIT primer sequences (3g): all designed first exon primers

PRIMER	NUMBER	SEQUENCE
Universal_5'start		GACGTGTGCTCTTCCGATCT
YAL012W	18	CGAACCCATTTCTTTGTCCA
YBR152W	19	AGAGCATCCAGACCAAACG
YDL137W	73	GAGAAGTCATGCAGAGAATGC
YDL189W	74	CCGTATCTTTATAATAGAGCTGGAA
YDR099W	81	GGTCAAGAAGATCAACAACAACA
YER102W	102	GAAGCTCACTACGGTCAAACC
YOL048C	127	CAAAGATATATCAAATATGGCTAAGCA
YIL009C-A	148	AAGAGATTCTCCTCACACCAAATACT
YIL123W	153	ACCCAATCTAGTGCTTCTTCTG
YKL150W	173	CAGCTGCTACCGCATTCTATT
YNL066W	228	GTCTTCGGCTCAAACAACCTC
YOR239W	243	GGAAGTTGTTGATGACTCTTGTCTT
YPL052W	246	GGATTTAAGGATTGGTGCG
YPL230W	256	GGTAACTTTGCCACCAATAGTGT

TABLE 24 *S. cerevisiae* SMIT primer sequences (4): intronless genes & terminal exons

PRIMER	NUMBER	SEQUENCE
Universal_5'start		GACGTGTGCTCTTCCGATCT
YBR111W-A	49	TCTTGTAGAATCAGGAAACTATGAAC
YDR424C	89	TGAGCGATGAAAATAAGAGTACG
YER074W-A	100	CAGTACTGAGCGAAGAAAGGTTC
YGL076C	117	GATCACAATGGCCGCTGA
YGR001C	126	ACTCCGACTCCGATTATGAA
YLR316C	191	AGATGTAAATACGCCCAAATAA
YPL198W	254	CAACGTCATAATGTCCACTGA

TABLE 25 *S. cerevisiae* SMIT primer sequences (5): first exon 2-intron genes

D.3 PRIMERS FOR *in vitro* TRANSCRIPTION AND TRANSCRIBED RNA SEQUENCES

In the course of developing and optimizing the SMIT and PacBio sequencing protocol *in vitro* transcription was used as defined template for ligation, RT and PCRs.

PRIMER	SEQUENCE
SMIT_RNA_P	AGUCGUAGCCUUAUCCGAGA- UUCAGCAAUA/3Phos/
SMIT_RNA	AGUCGUAGCCUUAUCCGAGA- UUCAGCAAUA
h18S 110 nt	GGAGAGAGAGAGAAUUACCC UCACUAAAGGGAGGAGAAGC UUAUCCCAAGAUCACUAC GAGCUUUUUAACUGCAGCAA CUUAAUAUACGCUAUUGGA GCUGGAAUUN
h18S 134 nt	GGAGAGAGAGAGAAUUACCC UCACUAAAGGGAGGAGAAGC UUAUCCCAAGAUCACUAC GAGCUUUUUAACUGCAGCAA CUUAAUAUACGCUAUUGGA GCUGGAAUUUCCGCGGCUGC UGUUCUAGAGGAUC
IVT_18ST7_fwd	GAAGAGAAGGAATTAATACGA- CTCA
IVT_18S_T7N_rev	NAATTCCAGCTCCAATAGCGTA
IVT_18S_T7A_rev	AAATTCCAGCTCCAATAGCGTA
IVT_18S_T7C_rev	CAATTCCAGCTCCAATAGCGTA
IVT_18S_T7G_rev	GAATTCCAGCTCCAATAGCGTA
IVT_18S_T7T_rev	TAATTCCAGCTCCAATAGCGTA
template (MegaShortScript 18S control)	GATTTAGGTGACACTATAGAA GAGAAGGAATTAATACGACTC ACTATAGGGAGAGAGAGAGAA TTACCCTCACTAAAGGGAGGA GAAGCTTATCCCAAGATCCAAC TACGAGCTTTTTAACTGCAGCA ACTTTAATATACGCTATTGGAGC TGGAATTTCCGCGGCTGCTGTT CTAGAGGATC

TABLE 26 3' end ligation (1): RNA sequences and IVT primer & template

PRIMER	SEQUENCE
spU1_fwd_SP6	CCAAGCCTTCATTTAGGTGACACTATAGAAGAGTGT CTTGGCATTGCACTGAGCCC
spU1_rev_T7	CAGAGATGCATAATACGACTCACTATAGGGAGAAAA- TTGCCCAAATGAGGGACGAAC
spU2_fwd_SP6	CCAAGCCTTCATTTAGGTGACACTATAGAAGAGCCT CTGGCTTGCTATGCTTTCCG
spU2_rev_T7	CAGAGATGCATAATACGACTCACTATAGGGAGAGCG TCGCTTGCCAGTAGTGC
spU3A_fwd_SP6	CCAAGCCTTCATTTAGGTGACACTATAGAAGAGTG TAATTTAAGAGCAGCTTCACCGCC
spU3A_rev_T7	CAGAGATGCATAATACGACTCACTATAGGGAGATGT CATCAAACGACACCACAGTTGTA
spU3B_fwd_SP6	CCAAGCCTTCATTTAGGTGACACTATAGAAGAGT GGCTGCTTTTGCAAAGCCAAGTG
spU3B_rev_T7	CAGAGATGCATAATACGACTCACTATAGGGAGAAC AGAAAACACGTCAGAAAACACCAGC
zfU2 191 nt	AUCGCUUCUCGGCCUUUUGGCUAAGAUCAAGUG UAGUAUCUGUUCUUAUCAGUUAAUAUCUGAUA CGUGCCCUACCCGGGCACCAUAUAUAAAUAUGA UUUUUGGAACAGGGAGAUGGAAUAGGGGCUUGC UCCGUCCACUCCACGUAUCGACCCGGUAUUGCAG UACAUCCGGGAACGGUGCACCCCU

TABLE 27 3' end ligation (2): RNA sequences and IVT primer & template

D.4 PRIMERS FOR PACBIO LIBRARIES

The nascent RNA PacBio library preparation involves 3' end adaptor ligation and a low cycle PCR after RT. Different experiments can be multiplexed using primers with different PacBio barcodes attached to the 3' end adaptor sequence in PCR. Those barcodes have been optimized in sequence with regard to the sequencing error profile of the technology.

PRIMER	SEQUENCE
3'end_DNA_adaptor (same as SMIT)	/5rApp/NNNNN- CTGTAGGCACCATCAAT/3ddC/
3'SMART_CDS_Primer_II_adap	AAGCAGTGGTATCAACGCAGA- GTACATTGATGGTGCCTACAG
SMARTer_II_A_Oligonucleotide (Clontech)	AAGCAGTGGTATCAACGCAGA- GTACXXXX
192_adap_BC1_rev	TCAGACGATGCGTCAT CATTGATGGTGCCTACAG
193_adap_BC2_rev	CTATACATGACTCTGC CATTGATGGTGCCTACAG
194_adap_BC3_rev	TACTAGAGTAGCACTC CATTGATGGTGCCTACAG
195_adap_BC4_rev	TGTGTATCAGTACATG CATTGATGGTGCCTACAG
196_adap_BC5_rev	ACACGCATGACACACT CATTGATGGTGCCTACAG
197_adap_BC6_rev	GATCTCTACTATATGC CATTGATGGTGCCTACAG
198_adap_BC7_rev	ACAGTCTATACTGCTG CATTGATGGTGCCTACAG
199_adap_BC8_rev	ATGATGTGCTACATCT CATTGATGGTGCCTACAG
200_adap_BC9_rev	CTGCGTGCTCTACGAC CATTGATGGTGCCTACAG
201_adap_BC10_rev	GCGCGATACGATGACT CATTGATGGTGCCTACAG
202_adap_BC11_rev	CGCGCTCAGCTGATCG CATTGATGGTGCCTACAG
203_adap_BC12_rev	GCGCACGCACTACAGA CATTGATGGTGCCTACAG
204_adap_BC13_rev	AACTGACGTCGCGAC CATTGATGGTGCCTACAG
205_adap_BC14_rev	CGTCTATATACGTATA CATTGATGGTGCCTACAG

TABLE 28 PacBio library primers for nascent RNA library and gene-specific library with barcoded reverse primer (192-205)

BIBLIOGRAPHY

- Stuart Aitken, Ross D Alexander, and Jean D Beggs. Modelling reveals kinetic advantages of co-transcriptional splicing. *PLoS Computational Biology*, 7(10): e1002215, October 2011. (Cited on page 7.)
- Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, July 2006. (Cited on page 98.)
- Ross D Alexander, Steven A Innocente, J David Barrass, and Jean D Beggs. Splicing-dependent RNA polymerase pausing in yeast. *Molecular Cell*, 40(4): 582–593, November 2010. (Cited on pages 7 and 71.)
- Adam Ameur, Anna Wetterbom, Lars Feuk, and Ulf Gyllensten. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biology*, 11(3):R34, 2010. (Cited on page 10.)
- Adam Ameur, Ammar Zaghlool, Jonatan Halvardson, Anna Wetterbom, Ulf Gyllensten, Lucia Cavellier, and Lars Feuk. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural & Molecular Biology*, 18(12):1435–1440, November 2011. (Cited on pages 12, 20, 23, and 97.)
- Stephen Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, 9(13):3015–3027, July 1981. (Cited on page 7.)
- Joanna Andrecka, Robert Lewis, Florian Brückner, Elisabeth Lehmann, Patrick Cramer, and Jens Michaelis. Single-molecule tracking of mRNA exiting from RNA polymerase II. *Proceedings of the National Academy of Sciences*, 105(1):135–140, January 2008. (Cited on page 65.)
- L Aravind, Hidemi Watanabe, David J Lipman, and Eugene V Koonin. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21): 11319–11324, October 2000. (Cited on page 12.)
- Moritz Aschoff, Agnes Hotz-Wagenblatt, Karl-Heinz Glatting, Matthias Fischer, Roland Eils, and Rainer König. SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics*, 29(9):1141–1148, May 2013. (Cited on pages 10 and 76.)
- Kin Fai Au et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences*, 110(50):E4821–30, December 2013. (Cited on page 76.)
- Susan M Berget. Exon recognition in vertebrate splicing. *The Journal of Biological Chemistry*, 270(6):2411–2414, February 1995. (Cited on page 68.)

- Susan M Berget, Claire Moore, and Phillip A Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8):3171–3175, August 1977. (Cited on page 5.)
- Dev M Bhatt, Amy Pandya-Jones, Ann-Jay Tong, Iros Barozzi, Michelle M Lissner, Gioacchino Natoli, Douglas L Black, and Stephen T Smale. Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions. *Cell*, 150(2):279–290, July 2012. (Cited on pages 10, 19, 21, 22, 23, 73, 74, and 110.)
- Jaswant S Bhorjee and Thoru Pederson. Chromatin: its isolation from cultured mammalian cells with particular reference to contamination by nuclear ribonucleoprotein particles. *Biochemistry*, 12(14):2766–2773, July 1973. (Cited on page 10.)
- Alicia A Bicknell, Can Cenik, Hon N Chua, Frederick P Roth, and Melissa J Moore. Introns in UTRs: why we should stop ignoring them. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 34(12):1025–1034, December 2012. (Cited on page 71.)
- Nicole I Bieberstein, Fernando Carrillo Oesterreich, Korinna Straube, and Karla M Neugebauer. First exon length controls active chromatin signatures and transcription. *CellReports*, 2(1):62–68, July 2012. (Cited on pages 6 and 71.)
- Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development*, 23(12):1379–1386, June 2009. (Cited on page 9.)
- David Botstein and Gerald R Fink. Yeast: an experimental organism for 21st Century biology. *Genetics*, 189(3):695–704, November 2011. (Cited on page 15.)
- Hannes Braberg et al. From Structure to Systems: High-Resolution, Quantitative Genetic Analysis of RNA Polymerase II. *Cell*, 154(4):775–788, August 2013. (Cited on page 6.)
- Ralph L Brinster, James M Allen, Richard R Behringer, Richard E Gelinas, and Richard D Palmiter. Introns increase transcriptional efficiency in transgenic mice. *Proceedings of the National Academy of Sciences*, 85(3):836–840, February 1988. (Cited on page 6.)
- Daria Brinzevich, George R Young, Robert Sebra, Juan Ayllon, Susan M Maio, Gintaras Deikus, Benjamin K Chen, Ana Fernandez-Sesma, Viviana Simon, and Lubbertus C F Mulder. HIV-1 interacts with human endogenous retrovirus K (HML-2) envelopes derived from human primary lymphocytes. *Journal of virology*, 88(11):6213–6223, June 2014. (Cited on page 76.)
- Mattia Brugiolo, Lydia Herzal, and Karla M Neugebauer. Counting on co-transcriptional splicing. *F1000prime reports*, 5(9):9, 2013. (Cited on pages 6, 10, 24, 65, 71, and 73.)

- Fernando Carrillo Oesterreich, Stephan Preibisch, and Karla M Neugebauer. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Molecular Cell*, 40(4):571–581, November 2010. (Cited on pages 6, 10, 24, 26, 27, 66, 68, 69, 70, 71, 73, 74, 79, 80, 81, 83, 89, and 134.)
- Fernando Carrillo Oesterreich, Nicole Bieberstein, and Karla M Neugebauer. Pause locally, splice globally. *Trends in Cell Biology*, 21(6):328–335, June 2011. (Cited on page 111.)
- Can Cenik, Adnan Derti, Joseph C Mellor, Gabriel F Berriz, and Frederick P Roth. Genome-wide functional analysis of human 5' untranslated region introns. *Genome Biology*, 11(3):R29, 2010. (Cited on page 71.)
- Keerthi T Chathoth, J David Barrass, Shaun Webb, and Jean D Beggs. A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation. *Molecular Cell*, 53(5):779–790, March 2014. (Cited on page 80.)
- Chen-Shan Chin et al. The origin of the Haitian cholera outbreak strain. *The New England journal of medicine*, 364(1):33–42, January 2011. (Cited on page 9.)
- Louise T Chow, Richard E Gelinis, Thomas R Broker, and Richard J Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, September 1977. (Cited on page 5.)
- L Stirling Churchman and Jonathan S Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373, January 2011. (Cited on pages 44, 53, 54, 66, and 137.)
- L Stirling Churchman and Jonathan S Weissman. *Native Elongating Transcript Sequencing (NET-seq)*. John Wiley & Sons, Inc., Hoboken, NJ, USA, May 2012. (Cited on pages 88 and 96.)
- Charles Cooke, Holly Hans, and James C Alwine. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. *Molecular and Cellular Biology*, 19(7):4971–4979, July 1999. (Cited on pages 70 and 71.)
- Antoine Coulon, Matthew L Ferguson, Valeria de Turris, Murali Palangat, Carson C Chow, and Daniel R Larson. Author response. *eLife*, 3:1–22, October 2014. (Cited on page 7.)
- Victoria H Cowling. Regulation of mRNA cap methylation. *The Biochemical journal*, 425(2):295–302, January 2010. (Cited on page 5.)
- Patrick Cramer. RNA polymerase II structure: from core to functional complexes. *Current opinion in genetics & development*, 14(2):218–226, April 2004. (Cited on page 3.)
- Paula Cramer, C Gustavo Pesce, Francisco E Baralle, and Alberto R Kornblihtt. Functional association between promoter structure and transcript alternative splicing. *Proceedings of the National Academy of Sciences of the United States of America*, 94(21):11456–11460, October 1997. (Cited on pages 71 and 73.)

- Christian Kroun Damgaard, Søren Kahns, Søren Lykke-Andersen, Anders Lade Nielsen, Torben Heick Jensen, and Jørgen Kjems. A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. *Molecular Cell*, 29(2):271–278, February 2008. (Cited on page 70.)
- Sérgio Fernandes de Almeida et al. Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nature Structural & Molecular Biology*, 18(9):977–983, September 2011. (Cited on page 6.)
- Eleonora de Klerk and Peter A C 't Hoen. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in genetics : TIG*, 31(3):128–139, March 2015. (Cited on page 7.)
- Manuel de la Mata, Celina Lafaille, and Alberto R Kornblihtt. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA*, 16(5):904–912, May 2010. (Cited on page 7.)
- Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013. (Cited on page 10.)
- Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105–e105, September 2008. (Cited on page 8.)
- Gwendal Dujardin et al. Transcriptional elongation and alternative splicing. *Biochimica et biophysica acta*, 1829(1):134–140, January 2013. (Cited on page 71.)
- Gwendal Dujardin, Celina Lafaille, Manuel de la Mata, Luciano E Marasco, Manuel J Munoz, Catherine Le Jossic-Corcós, Laurent Corcos, and Alberto R Kornblihtt. How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Molecular Cell*, 54(4):683–690, May 2014. (Cited on page 6.)
- Arik Dvir. Promoter escape by RNA polymerase II. *Biochimica et biophysica acta*, 1577(2):208–223, September 2002. (Cited on page 3.)
- Michael J Dye and Nick J Proudfoot. Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. *Molecular Cell*, 3(3):371–378, March 1999. (Cited on pages 70 and 71.)
- John Eid et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, January 2009. (Cited on pages 9 and 76.)
- Adam C English et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, 7(11):e47768, 2012. (Cited on page 76.)
- Gerald R Fink. Pseudogenes in yeast? *Cell*, 49(1):5–6, April 1987. (Cited on pages 12 and 119.)

- Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6):461–465, June 2010. (Cited on page 9.)
- Nova Fong, Hyunmin Kim, Yu Zhou, Xiong Ji, Jinsong Qiu, Tassa Saldi, Katrina Diener, Ken Jones, Xiang-Dong Fu, and David L Bentley. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes & Development*, 28(23):2663–2676, December 2014. (Cited on page 6.)
- Mathieu Foquet, Kevan T Samiee, Xiangxu Kong, Bidhan P Chauduri, Paul Lundquist, Stephen W Turner, Jake Freudenthal, and Daniel B Roitman. Improved fabrication of zero-mode waveguides for single-molecule detection. *Journal of Applied Physics*, 103(3):034301, February 2008. (Cited on page 9.)
- Susan L Forsburg. The yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*: models for cell biology research. *Gravitational and space biology bulletin : publication of the American Society for Gravitational and Space Biology*, 18(2):3–9, June 2005. (Cited on pages 12 and 68.)
- Kristi L Fox-Walsh, Yimeng Dou, Bianca J Lam, She-Pin Hung, Pierre F Baldi, and Klemens J Hertel. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16176–16181, November 2005. (Cited on pages 68 and 73.)
- Janina Görnemann, Kimberly M Kotovic, Katja Hujer, and Karla M Neugebauer. Cotranscriptional Spliceosome Assembly Occurs in a Stepwise Fashion and Requires the Cap Binding Complex. *Molecular Cell*, 19(1):53–63, July 2005. (Cited on pages 5, 7, 66, and 70.)
- Cameron J Grisdale, Lisa C Bowers, Elizabeth S Didier, and Naomi M Fast. Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote. *BMC Genomics*, 14(1):207, 2013. (Cited on pages 12 and 22.)
- Ingrid Grummt. Regulation of mammalian ribosomal gene transcription by RNA polymerase I. *Progress in nucleic acid research and molecular biology*, 62:109–154, 1999. (Cited on page 3.)
- Muxin Gu, Yanin Naiyachit, Thomas J Wood, and Catherine B Millar. H2A.Z marks antisense promoters and has positive effects on antisense transcript levels in budding yeast. *BMC Genomics*, 16(1):99, 2015. (Cited on pages 96 and 120.)
- Gal Haimovich, Daniel A Medina, Sebastien Z Causse, Manuel Garber, Gonzalo Millán-Zambrano, Oren Barkai, Sebastián Chávez, José E Pérez-Ortín, Xavier Darzacq, and Mordechai Choder. Gene expression is circular: factors for mRNA degradation also foster mRNA synthesis. *Cell*, 153(5):1000–1011, May 2013. (Cited on page 7.)

- S Blair Hedges. The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838–849, November 2002. (Cited on page 13.)
- Lydia Herzal and Karla M Neugebauer. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods (San Diego, Calif.)*, April 2015. (Cited on pages 16, 65, 73, and 97.)
- Patricia Heyn, Martin Kircher, Andreas Dahl, Janet Kelso, Pavel Tomancak, Alex T Kalinka, and Karla M Neugebauer. The earliest transcribed zygotic genes are short, newly evolved, and different across species. *CellReports*, 6(2):285–292, January 2014. (Cited on page 10.)
- Steve Hoffmann et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biology*, 15(2):R34, 2014. (Cited on pages 10, 41, 43, and 97.)
- Jing-Ping Hsin and James L Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & Development*, 26(19):2119–2137, October 2012. (Cited on page 3.)
- Martina Huranová, Ivan Ivani, Ales Benda, Ina Poser, Yehuda Brody, Martin Hof, Yaron Shav-Tal, Karla M Neugebauer, and David Staněk. The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. *The Journal of Cell Biology*, 191(1):75–86, October 2010. (Cited on pages 6 and 7.)
- Daniel C Jeffares, Christopher J Penkett, and J Bahler. Rapidly regulated genes are intron poor. *Trends in Genetics*, pages 1–4, July 2008. (Cited on page 71.)
- Jill M Johnsen, Deborah A Nickerson, and Alex P Reiner. Massively parallel sequencing: the new frontier of hematologic genomics. *Blood*, 122(19):3268–3275, November 2013. (Cited on page 8.)
- Iris Jonkers and John T Lis. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3):167–177, March 2015. (Cited on page 66.)
- Iris Jonkers, Hojoong Kwak, and John T Lis. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, 3:1–25, April 2014. (Cited on pages 65, 66, and 71.)
- Norbert Kaeufer and Judith Potashkin. Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic Acids Research*, pages 1–8, July 2000. (Cited on pages 14 and 68.)
- Daisuke Kaida, Michael G Berg, Ihab Younis, Mumtaz Kasim, Larry N Singh, Lili Wan, and Gideon Dreyfuss. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, 468(7324):664–668, December 2010. (Cited on page 70.)

- Yarden Katz, Eric T Wang, Edoardo M Airoidi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, December 2010. (Cited on pages 10 and 76.)
- Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624, April 2004. (Cited on page 118.)
- W James Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, April 2002. (Cited on page 76.)
- Yevgenia L Khodor, Joseph Rodriguez, Katharine C Abruzzi, Chih-Hang H A Tang, Michael T Marr, and Michael Rosbash. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes & Development*, 25(23):2502–2512, December 2011. (Cited on pages 7, 10, 11, 24, 68, 69, 70, 73, 74, 97, and 105.)
- Yevgenia L Khodor, Jerome S Menet, Michael Tolan, and Michael Rosbash. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA*, 18(12):2174–2186, December 2012. (Cited on pages 7, 10, 19, 20, 21, 22, 23, 69, 73, 74, 105, and 110.)
- Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013. (Cited on pages 10, 19, 26, and 97.)
- Martin Kircher, Patricia Heyn, and Janet Kelso. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, 12(1):382, 2011. (Cited on page 9.)
- Claudia L Kleinman et al. Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR. *Nature Genetics*, 46(1):39–44, January 2014. (Cited on page 76.)
- Shihoko Kojima, Elaine L Sher-Chen, and Carla B Green. Circadian control of mRNA polyadenylation dynamics regulates rhythmic protein expression. *Genes & Development*, 26(24):2724–2736, December 2012. (Cited on page 74.)
- Jonas Korlach, Keith P Bjornson, Bidhan P Chaudhuri, Ronald L Cicero, Benjamin A Flusberg, Jeremy J Gray, David Holden, Ravi Saxena, Jeffrey Wegener, and Stephen W Turner. Real-time DNA sequencing from single polymerase molecules. *Methods in enzymology*, 472:431–455, 2010. (Cited on pages 9 and 76.)
- Roger D Kornberg. Eukaryotic transcriptional control. *Trends in Cell Biology*, 9(12):M46–9, December 1999. (Cited on page 3.)
- Alberto R Kornblihtt. Transcriptional control of alternative splicing along time: ideas change, experiments remain. *RNA*, 21(4):670–672, April 2015. (Cited on page 7.)

- Kristin S Koutmou, Anthony P Schuller, Julie L Brunelle, Aditya Radhakrishnan, Sergej Djuranovic, and Rachel Green. Ribosomes slide on lysine-encoding homopolymeric A stretches. *eLife*, 4, 2015. (Cited on page 74.)
- Jason N Kuehner, Erika L Pearson, and Claire Moore. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews Molecular Cell Biology*, 12(5):283–294, May 2011. (Cited on page 5.)
- Andreas N Kuhn and Norbert F Käufer. Pre-mRNA splicing in *Schizosaccharomyces pombe*: regulatory role of a kinase conserved from fission yeast to mammals. *Current Genetics*, 42(5):241–251, February 2003. (Cited on page 14.)
- Doris Kupfer, Scott Drabenstot, Kent Buchanan, Hongshing Lai, Hua Zhu, David Dyer, Bruce Roe, and Juneann Murphy. Introns and Splicing Elements of Five Diverse Fungi. *Eukaryotic cell*, pages 1–13, September 2004. (Cited on page 14.)
- Hojoong Kwak, Nicholas J Fuda, Leighton J Core, and John T Lis. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science*, 339(6122):950–953, February 2013. (Cited on pages 3 and 66.)
- Kon Yew Kwek, Shona Murphy, Andre Furger, Benjamin Thomas, William O’Gorman, Hiroshi Kimura, Nick J Proudfoot, and Alexandre Akoulitchev. U1 snRNA associates with TFIID and regulates transcriptional initiation. *Nature structural biology*, 9(11):800–805, November 2002. (Cited on page 70.)
- Scott A Lacadie and Michael Rosbash. Cotranscriptional spliceosome assembly dynamics and the role of U1 snRNA:5’ss base pairing in yeast. *Molecular Cell*, 19(1):65–75, July 2005. (Cited on pages 5, 7, and 66.)
- Scott A Lacadie, Daniel F Tardiff, Sebastian Kadener, and Michael Rosbash. In vivo commitment to yeast cotranscriptional splicing is sensitive to transcription elongation mutants. *Genes & Development*, 20(15):2055–2066, August 2006. (Cited on pages 5, 6, and 7.)
- Daniel H Lackner, Beilharz T, Samuel Marguerat, Juan Mata, Stephen Watt, Falk Schubert, Thomas Preiss, and J Bahler. A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. *Molecular Cell*, pages 1–15, March 2007. (Cited on page 71.)
- E S Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. (Cited on page 7.)
- Peter A Larsen and Timothy P L Smith. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC immunology*, 13(1):52, 2012. (Cited on page 76.)
- Nelson C Lau, Lee P Lim, Earl G Weinstein, and David P Bartel. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, pages 1–5, October 2001. (Cited on page 54.)

- Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, 7(9):709–715, August 2010. (Cited on pages 9 and 80.)
- Bing Li, Michael Carey, and Jerry L Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, February 2007. (Cited on page 5.)
- Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. (Cited on page 97.)
- Hua Li, Jingyi Hou, Ling Bai, Chuansheng Hu, Pan Tong, Yani Kang, Xiaodong Zhao, and Zhifeng Shao. Genome-wide analysis of core promoter structures in *Schizosaccharomyces pombe* with DeepCAGE. *RNA Biology*, 12(5):525–537, May 2015. (Cited on page 70.)
- Xialu Li and James L Manley. Cotranscriptional processes and their influence on genome stability. *Genes & Development*, 20(14):1838–1847, July 2006. (Cited on page 4.)
- Imke Listerman, Aparna K Sapra, and Karla M Neugebauer. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nature Structural & Molecular Biology*, 13(9):815–822, September 2006. (Cited on page 5.)
- Alessandro Marcello. RNA polymerase II transcription on the fast lane. *Transcription*, 3(1):29–34, October 2014. (Cited on page 6.)
- Elaine R Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008. (Cited on page 8.)
- M Martin. *Cutadapt removes adapter sequences from high-throughput sequencing reads*. *EMBnet. journal 17 h ttp*. [journal.embnet.org/index.php/embnetjournal/article/...](http://journal.embnet.org/index.php/embnetjournal/article/), 2014. (Cited on page 99.)
- Robert M Martin, José Rino, Célia Carvalho, Tomas Kirchhausen, and Maria Carmo-Fonseca. Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. *CellReports*, 4(6):1144–1155, September 2013. (Cited on pages 6 and 7.)
- Fuensanta W Martinez-Rucobo, Rebecca Kohler, Michiel van de Waterbeemd, Albert J R Heck, Matthias Hemann, Franz Herzog, Holger Stark, and Patrick Cramer. Molecular Basis of Transcription-Coupled Pre-mRNA Capping. *Molecular Cell*, May 2015. (Cited on pages 5, 6, 65, and 83.)
- Sandra Bento Martins, José Rino, Teresa Carvalho, Célia Carvalho, Minoru Yoshida, Jasmim Mona Klose, Sérgio Fernandes de Almeida, and Maria Carmo-Fonseca. Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3' end of human genes. *Nature Structural & Molecular Biology*, 18(10):1115–1123, September 2011. (Cited on page 70.)

- Paul B Mason and Kevin Struhl. Distinction and Relationship between Elongation Rate and Processivity of RNA Polymerase II In Vivo. *Molecular Cell*, 17(6): 831–840, March 2005. (Cited on pages 6 and 65.)
- Andreas Mayer, Michael Lidschreiber, Matthias Siebert, Kristin Leike, Johannes Söding, and Patrick Cramer. Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology*, 17(10):1272–1278, September 2010. (Cited on pages 3 and 66.)
- Andreas Mayer, Martin Heidemann, Michael Lidschreiber, Amelie Schrieck, Mai Sun, Corinna Hintermair, Elisabeth Kremmer, Dirk Eick, and Patrick Cramer. CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science*, 336(6089):1723–1725, June 2012. (Cited on pages 3, 66, and 67.)
- Dominik M Meinel, Cornelia Burkert-Kautzsch, Anja Kieser, Eoghan O’Duibhir, Matthias Siebert, Andreas Mayer, Patrick Cramer, Johannes Söding, Frank C P Holstege, and Katja Sträßler. Recruitment of TREX to the transcription machinery by its direct binding to the phospho-CTD of RNA polymerase II. *PLoS Genetics*, 9(11):e1003914, November 2013. (Cited on page 66.)
- Tim R Mercer, Michael B Clark, Joanna Crawford, Marion E Brunck, Daniel J Gerhardt, Ryan J Taft, Lars K Nielsen, Marcel E Dinger, and John S Mattick. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature Protocols*, 9(5):989–1009, May 2014. (Cited on page 80.)
- Christian Miller et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular systems biology*, 7(1):458–458, January 2011. (Cited on page 10.)
- Shin Moteki and David Price. Functional coupling of capping and transcription of mRNA. *Molecular Cell*, 10(3):599–609, September 2002. (Cited on page 6.)
- Georgette Moyle-Heyrman, Tetiana Zaichuk, Liqun Xi, Quanwei Zhang, Olke C Uhlenbeck, Robert Holmgren, Jonathan Widom, and Ji-Ping Wang. Chemical map of *Schizosaccharomyces pombe* reveals species-specific features in nucleosome positioning. *Proceedings of the National Academy of Sciences*, 110(50): 20158–20163, December 2013. (Cited on page 66.)
- Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, June 2008. (Cited on page 80.)
- Nagarjuna Nagaraj, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, and Matthias Mann. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7(1):548–548, 2011. (Cited on pages 98 and 120.)

- Fernando Carrillo Oesterreich, Stephan Preibisch, and Karla M Neugebauer. Global Analysis of Nascent RNA Reveals Transcriptional Pausing in Terminal Exons. *Molecular Cell*, 40(4):571–581, November 2010. (Cited on pages 10 and 53.)
- Yvonne N Osheim, Oscar L Miller, and Ann L Beyer. RNP particles at splice junction sequences on *Drosophila* chorion transcripts. *Cell*, 43(1):143–151, November 1985. (Cited on page 6.)
- Amy Pandya-Jones. Pre-mRNA splicing during transcription in the mammalian system. *Wiley Interdisciplinary Reviews: RNA*, 2(5):700–717, May 2011. (Cited on page 6.)
- Julie Parenteau et al. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Molecular biology of the cell*, 19(5):1932–1941, May 2008. (Cited on page 119.)
- Julie Parenteau, Mathieu Durand, Geneviève Morin, Jules Gagnon, Jean-François Lucier, Raymund J Wellinger, Benoit Chabot, and Sherif Abou Elela. Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell*, 147(2):320–331, October 2011. (Cited on pages 33 and 119.)
- Nick J Proudfoot. Ending the message: poly(A) signals then and now. *Genes & Development*, 25(17):1770–1782, September 2011. (Cited on page 5.)
- Irina Pulyakhina, Isabella Gazzoli, Peter-Bram 't Hoen, Nisha Verwey, Johan den Dunnen, Annemieke Aartsma-Rus, and Jeroen Laros. SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Research*, page gkv242, March 2015. (Cited on page 10.)
- Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, 2012. (Cited on pages 9 and 75.)
- Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. (Cited on page 97.)
- Charalampos Rallis, Sandra Codlin, and Jürg Bähler. TORC1 signaling inhibition by rapamycin and caffeine affect lifespan, global gene expression, and cell proliferation of fission yeast. *Aging cell*, 12(4):563–573, August 2013. (Cited on pages 34, 71, and 116.)
- Eric B Rasmussen and John T Lis. In vivo transcriptional pausing and cap formation on three *Drosophila* heat shock genes. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17):7923–7927, September 1993. (Cited on pages 65 and 83.)

- Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597, May 2015. (Cited on pages 7 and 8.)
- Donald C Rio, Manuel Ares, Gregory J Hannon, and Timothy W Nilsen. Removing Ribosomal RNAs. In *RNA A Laboratory Manual*. 2011. (Cited on page 77.)
- Barbara L Robberson, Gilbert J Cote, and Susan M Berget. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Molecular and Cellular Biology*, 10(1):84–94, January 1990. (Cited on page 68.)
- James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, January 2011. (Cited on page 97.)
- Charles M Romfo, Consuelo J Alvarez, Willem J van Heeckeren, Christopher J Webb, and Jo Ann Wise. Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Molecular and Cellular Biology*, 20(21):7955–7970, November 2000. (Cited on pages 68 and 73.)
- Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51, 2013. (Cited on pages 76 and 79.)
- Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutürk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, October 2011. (Cited on page 8.)
- F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977. (Cited on page 7.)
- Abbie Saunders, Leighton J Core, and John T Lis. Breaking barriers to transcription elongation. *Nature Reviews Molecular Cell Biology*, 7(8):557–567, August 2006. (Cited on page 3.)
- Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, March 2011. (Cited on page 100.)
- Dietmar Schreiner, Thi-Minh Nguyen, Giancarlo Russo, Steffen Heber, Andrea Patrignani, Erik Ahrné, and Peter Scheiffele. Targeted combinatorial alternative splicing generates brain region-specific repertoires of neurexins. *Neuron*, 84(2):386–398, October 2014. (Cited on page 76.)
- Schrage Schwartz, Eran Meshorer, and Gil Ast. Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology*, 16(9):990–995, September 2009. (Cited on page 66.)

- Ulrike Schwarze, Barbra J Starman, and Peter H Byers. Redefinition of exon 7 in the COL1A1 gene of type I collagen by an intron 8 splice-donor-site mutation in a form of osteogenesis imperfecta: influence of intron splice order on outcome of splice-site mutation. *American journal of human genetics*, 65(2):336–344, August 1999. (Cited on page 10.)
- Donald Sharon, Hagen Tilgner, Fabian Grubert, and Michael Snyder. A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, 31(11):1009–1014, November 2013. (Cited on pages 9 and 76.)
- Peter J Shepard and Klemens J Hertel. The SR protein family. *Genome Biology*, 10(10):242, 2009. (Cited on page 5.)
- Andrej Shevchenko, Henrik Tomas, Jan Havlis, Jesper V Olsen, and Matthias Mann. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols*, 1(6):2856–2860, 2006. (Cited on page 94.)
- Anna Shevchenko, Assen Roguev, Daniel Schaft, Luke Buchanan, Bianca Habermann, Cagri Sakalar, Henrik Thomas, Nevan J Krogan, Andrej Shevchenko, and A Francis Stewart. Chromatin Central: towards the comparative proteome by accurate mapping of the yeast proteomic environment. *Genome Biology*, 9(11):R167, 2008. (Cited on page 94.)
- Jarnail Singh and Richard A Padgett. Rates of in situ transcription and splicing in large human genes. *Nature Structural & Molecular Biology*, 16(11):1128–1133, October 2009. (Cited on page 7.)
- Ravi K Singh and Thomas A Cooper. Pre-mRNA splicing in disease and therapeutics. *Trends in molecular medicine*, 18(8):472–482, August 2012. (Cited on page 5.)
- Matthias Sipiczki. Where does fission yeast sit on the tree of life? *Genome Biology*, pages 1–4, August 2000. (Cited on pages 12 and 13.)
- Henrik Stranneheim and Joakim Lundeberg. Stepping stones in DNA sequencing. *Biotechnology journal*, 7(9):1063–1073, September 2012. (Cited on page 9.)
- Mai Sun, Björn Schwalb, Daniel Schulz, Nicole Pirkl, Stefanie Etzold, Laurent Larivière, Kerstin C Maier, Martin Seizl, Achim Tresch, and Patrick Cramer. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Research*, 22(7):1350–1359, July 2012. (Cited on page 71.)
- Daniel F Tardiff and Michael Rosbash. Arrested yeast splicing complexes indicate stepwise snRNP recruitment during in vivo spliceosome assembly. *RNA*, 12(6):968–979, April 2006. (Cited on pages 6 and 7.)
- Daniel F Tardiff, Scott A Lacadie, and Michael Rosbash. A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Molecular Cell*, 24(6):917–929, December 2006. (Cited on page 66.)

- Sean Thomas, Jason G Underwood, Elizabeth Tseng, Alisha K Holloway, and Bench To Basinet CvDC Informatics Subcommittee. Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLoS ONE*, 9(4):e94650, 2014. (Cited on page 76.)
- Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, March 2013. (Cited on page 97.)
- Hagen Tilgner, Christoforos Nikolaou, Sonja Althammer, Michael Sammeth, Miguel Beato, Juan Valcárcel, and Roderic Guigo. Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology*, 16(9):996–1001, September 2009. (Cited on page 66.)
- Hagen Tilgner, David G Knowles, Rory Johnson, Carrie A Davis, Sudipto Chakraborty, Sarah Djebali, Joao Curado, Michael Snyder, Thomas R Gingeras, and Roderic Guigo. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, September 2012. (Cited on pages 7, 10, 12, 69, 73, and 97.)
- Hagen Tilgner, Fabian Grubert, Donald Sharon, and Michael P Snyder. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences*, 111(27):9869–9874, July 2014. (Cited on page 76.)
- Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012. (Cited on pages 10 and 76.)
- Kevin J Travers, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, 38(15):e159–e159, August 2010. (Cited on page 9.)
- Barbara Treutlein, Ozgun Gokce, Stephen R Quake, and Thomas C Südhof. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proceedings of the National Academy of Sciences*, 111(13):E1291–9, April 2014. (Cited on page 76.)
- Diana Y Vargas, Khyati Shah, Mona Batish, Michael Levandoski, Sourav Sinha, Salvatore A E Marras, Paul Schedl, and Sanjay Tyagi. Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing. *Cell*, 147(5):1054–1065, November 2011. (Cited on page 7.)
- Sebastien Viollet, Ryan T Fuchs, Daniela B Munafo, Fanglei Zhuang, and Gregory B Robb. T4 RNA Ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotechnology*, 11(1):72, July 2011. (Cited on page 54.)

- Kristoffer Vitting-Seerup, Bo Torben Porse, Albin Sandelin, and Johannes Waage. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC bioinformatics*, 15(1):81, 2014. (Cited on pages 10 and 76.)
- Markus C Wahl, Cindy L Will, and Reinhard Luhrmann. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, February 2009. (Cited on pages 6 and 12.)
- Kai Wang et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178–e178, October 2010. (Cited on page 10.)
- Peter L Wang, Yun Bao, Muh-Ching Yee, Steven P Barrett, Gregory J Hogan, Mari N Olsen, José R Dinneny, Patrick O Brown, and Julia Salzman. Circular RNA is expressed across the eukaryotic tree of life. *PLoS ONE*, 9(6):e90859, 2014. (Cited on page 42.)
- Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, June 2008. (Cited on page 73.)
- Ian M Willis. RNA polymerase III. Genes, factors and transcriptional specificity. *European journal of biochemistry / FEBS*, 212(1):1–11, February 1993. (Cited on page 3.)
- Lukas Windhager et al. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research*, 22(10):2031–2042, October 2012. (Cited on page 10.)
- V Wood et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415(6874):871–880, February 2002. (Cited on page 87.)
- Thomas D Wu and Colin K Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, May 2005. (Cited on pages 76 and 99.)
- Jerome Wuarin and Ueli Schibler. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Molecular and Cellular Biology*, 14(11):7219–7225, November 1994. (Cited on page 10.)
- Jun Yu, Zhiyong Yang, Miho Kibukawa, Marcia Paddock, Douglas A Passey, and Gane Ka-Shu Wong. Minimal introns are not "junk". *Genome Research*, 12(8):1185–1189, August 2002. (Cited on page 70.)
- Ammar Zaghlool, Adam Ameer, Linnea Nyberg, Jonatan Halvardson, Manfred Grabherr, Lucia Cavellier, and Lars Feuk. Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. *BMC Biotechnology*, 13(1):99, 2013. (Cited on page 10.)

- Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, December 2008. (Cited on page [65](#).)
- Wei Zhang, Paul Ciclitira, and Joachim Messing. PacBio sequencing of gene families - a case study with wheat gluten genes. *Gene*, 533(2):541–546, January 2014. (Cited on page [76](#).)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both \LaTeX and \LyX :

<http://code.google.com/p/classicthesis/>

DECLARATION

Declaration according to §5.5 of the doctorate regulations

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other German or foreign examination board.

The thesis work was conducted from 06/01/2011 to 05/31/2015 under the supervision of Prof. Karla M. Neugebauer, Ph.D. at the Max Planck Institute of Molecular Cell Biology and Genetics and Yale University.

I declare that I have not undertaken any previous unsuccessful doctorate proceedings.

I declare that I recognize the doctorate regulations of the Fakultät für Mathematik und Naturwissenschaften of the Technische Universität Dresden.

Erklärung entsprechend §5.5 der Promotionsordnung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die Dissertation wurde im Zeitraum vom 01.06.2011 bis 31.05.2015 verfasst und von Prof. Karla M. Neugebauer, Ph.D., Max Planck Institut für Molekulare Zellbiologie und Genetik und Yale University, betreut.

Meine Person betreffend erkläre ich hiermit, dass keine früheren erfolglosen Promotionsverfahren stattgefunden haben.

Ich erkenne die Promotionsordnung der Fakultät für Mathematik und Naturwissenschaften, Technische Universität Dresden an.

Dresden, June 2015

Lydia Herzel