UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Prevalence and Significance of Nonsense Mediated mRNA Decay Coupled with Alternative Splicing in Diverse Eukaryotic Organisms

Permalink
https://escholarship.org/uc/item/3808v7d8

Author
French, Courtney Elizabeth

Publication Date
2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Prevalence and Significance of Nonsense-Mediated mRNA Decay Coupled with Alternative Splicing in Diverse Eukaryotic Organisms

By

Courtney Elizabeth French

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Steven E. Brenner, Co-Chair
Professor Donald C. Rio, Co-Chair
Professor Britt A. Glaunsinger
Professor Sandrine Dudoit

Spring 2016

Prevalence and Significance of Nonsense-Mediated mRNA Decay Coupled with Alternative Splicing in Diverse Eukaryotic Organisms

Copyright 2016
by
Courtney Elizabeth French

Abstract
Prevalence and Significance of Nonsense-Mediated mRNA Decay Coupled with Alternative Splicing in Diverse Eukaryotic Organisms

by

Courtney Elizabeth French

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Steven E. Brenner, Co-Chair

Professor Donald C. Rio, Co-Chair

Alternative splicing plays a crucial role in increasing the amount of protein diversity and in regulating gene expression at the post-transcriptional level. In humans, almost all genes produce more than one mRNA isoform and, while the fraction varies, many other species also have a substantial number of alternatively spliced genes. Alternative splicing is regulated by splicing factors, often in a developmental time- or tissue-specific manner. Mis-regulation of alternative splicing, via mutations in splice sites, splicing regulatory elements, or splicing factors, can lead to disease states, including cancers. Thus, characterizing how alternative splicing shapes the transcriptome will lead to greater insights into the regulation of numerous cellular pathways and many aspects of human health.

A critical tool for investigating alternative splicing is high-throughput mRNA sequencing (RNA-seq). This technology produces hundreds of millions of short (~100bp) sequencing reads from mRNA molecules and can be used to both discover novel transcripts and to quantify the expression of transcripts. While short read length is a limitation of the technology in its current form, RNA-seq has resulted in the discovery of hundreds of thousands of new transcripts and revealed an increased complexity of the transcriptome via alternative splicing, particularly in human. Here, I used RNA-seq analysis to investigate the global effect of post-transcriptional regulation via alternative splicing coupled to nonsense-mediated mRNA decay and to examine natural human variation in alternative splicing, particularly in genes associated with differential therapeutic drug response.

The nonsense-mediated mRNA decay pathway (NMD), which degrades transcripts containing a premature termination codon, plays an important role in post-transcriptional gene regulation when coupled to alternative splicing. If a gene produces an alternative isoform that is targeted by NMD, the mRNA abundance of the protein-producing transcripts can be post-transcriptionally regulated at the alternative splicing level. This has been shown to be important in the regulation of a number of genes, including many of the splicing factors themselves. I have used RNA-seq analysis on cells where NMD has been

inhibited to discover alternative isoforms that are NMD targets on a genome-wide scale in human and a number of diverse other eukaryotic species. I found that around 20% of expressed human genes are potentially regulated by alternative splicing coupled to NMD and that they fall into many different functional categories. I also found that hundreds to thousands of genes produce NMD-targeted alternative isoforms in each of frog, zebrafish, fly, fission yeast, and plant, highlighting the prevalence of this relatively under-studied method of gene regulation across the three major branches of eukaryotic organisms. I also gained insight into the features that define NMD targets, which are thought to vary between species although the field is still unclear. I find that an exon-exon junction downstream of the termination codon is a much stronger predictor of NMD than 3' UTR length in every species except yeast.

I also used RNA-seq to investigate alternative splicing in genes of pharmacologic importance. Natural human variation in the expression level and activity of genes involved in drug disposition and action ("pharmacogenes") can affect drug response and toxicity. Previous studies have relied primarily on microarrays to understand gene expression differences, or have focused on a single tissue or small number of samples. Here, we used RNA-seq to determine the expression levels and alternative splicing of 389 selected pharmacogenes across four pharmacologically relevant tissues (liver, kidney, heart and adipose) and lymphoblastoid cell lines (LCLs), which are used widely in pharmacogenomics studies. Analysis of data from 18 different individuals for each of the 5 tissues (90 samples in total) revealed substantial variation in both expression levels and splicing across samples and tissue types. Comparison with an independent RNA-seq dataset yielded a consistent picture. This in-depth exploration also revealed 183 splicing events in pharmacogenes that were previously not annotated. Overall, this study serves as a rich resource for the research community to inform biomarker and drug discovery and use.[*]

In conclusion, the roles of alternative splicing and NMD in the regulation of cellular processes and in human health are wide-open but critical fields of study. Advancements in sequencing technologies have had and will continue to have a huge impact on the studies of these mechanisms. New long-read technologies will likely soon be readily available and promise to greatly increase our ability to accurately interpret RNA-seq results. As the cost of sequencing continues to decrease, more and more data will be generated, allowing for a better view of how the transcriptome varies between individuals and shapes differential disease risks and drug responses.

[*] This paragraph was co-written with Aparna Chhibber and Sook Wah Yee and modified from a previously published work: Chhibber A*, French CE*, Yee SW*, Gamazon ER*, et al. 2016. Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *Pharmacogenomics Journal.* *co-first authors

Table of Contents

CHAPTER 1 USING RNA-SEQ ANALYSIS FOR ISOFORM EXPRESSION AND ALTERNATIVE SPLICING INVESTIGATIONS 1
ABSTRACT 1
INTRODUCTION TO RNA-SEQ 1
READ ALIGNMENT 2
ASSEMBLY OF NOVEL TRANSCRIPTS 2
QUANTIFYING TRANSCRIPT EXPRESSION 3
USING A REFERENCE GENE ANNOTATION 6
IMPACT OF READ DEPTH 7
READ DEPTH VERSUS NUMBER OF SAMPLES 9
CONCLUSIONS 11
REFERENCES 12

CHAPTER 2 TRANSCRIPTOME ANALYSIS OF ALTERNATIVE SPLICING COUPLED WITH NONSENSE MEDIATED MRNA DECAY IN HUMAN CELLS 15
ABSTRACT 15
INTRODUCTION 15
RESULTS 17
Over two thousand genes produce transcripts identified as confident NMD targets 17
Diverse categories of genes are affected by NMD, particularly splicing 21
NMD-targeted genes are enriched for ultraconserved elements 25
The 50nt Rule is a strong predictor of NMD while a longer 3' UTR has little effect 25
Premature termination codons generated by alternative splicing events or uORFS 27
DISCUSSION 28
MATERIALS AND METHODS 30
REFERENCES 32

CHAPTER 3 EVIDENCE FOR PERVASIVE REGULATION VIA NONSENSE-MEDIATED MRNA DECAY ACROSS SEVERAL DIVERSE SPECIES 38
ABSTRACT 38
INTRODUCTION 38
RESULTS 40
Analysis of RNA-seq data from UPF1-depletion studies for seven eukaryotic species 40
Features of NMD-affected transcripts 42
Pervasiveness of NMD-degraded isoforms and potential for regulation 47
DISCUSSION 47
MATERIALS AND METHODS 51
REFERENCES 52

CHAPTER 4 TRANSCRIPTOMIC VARIATION OF PHARMACOGENES IN MULTIPLE HUMAN TISSUES AND LYMPHOBLASTOID CELL LINES 56
ABSTRACT 56
INTRODUCTION 56
RESULTS 57

The Pharmacogenomics Research Network RNA-Seq Project *57*
Analysis of PGRN pharmacogene gene expression *58*
Analysis of PGRN pharmacogene splicing *62*
DISCUSSION 65
MATERIALS AND METHODS 70
REFERENCES 72

List of Figures

Figure 1.1 Unsupported novel junctions in Cufflinks-assembled transcripts 4
Figure 1.2 Using a Shannon entropy score to detect false positive junctions 5
Figure 1.3 Transcript quantification depends on fragment depth 10
Figure 1.4 Increasing fragment depth decreases transcript quantification uncertainty 10
Figure 1.5 The effect of number of samples and read depth on observing splicing events 11
Figure 2.1 Transcripts with a PTC_{50nt} are more likely to increase when NMD is inhibited 19
Figure 2.2 Distribution of expression levels of NMD targets and non-NMD targets 20
Figure 2.3 qPCR validation of NMD targets 21
Figure 2.4 Known and novel splicing events resulting in NMD targets 22
Figure 2.5 NMD-targeted genes fall into diverse functional categories 24
Figure 2.6 The 50nt Rule is a strong predictor of NMD and a longer 3' UTR is not 26
Figure 3.1 Effect of the 50nt Rule on transcript stability 43
Figure 3.2 Effect of 3' UTR length on transcript stability 45
Figure 3.3 Alternative splicing coupled to NMD in seven species 48
Figure 3.4 NMD-targeted genes in seven species 49
Figure 4.1 Overview of the Pharmacogenomics Research Network RNA-seq project 58
Figure 4.2 Heatmap of the 389 PGRN pharmacogenes' expression across 90 samples 60
Figure 4.3 Gene expression by sample across each tissue for selected genes 61
Figure 4.4 Tissue-specific and/or novel splice events in pharmacogenes 66
Figure 4.5 Novel junctions have lower PSI values and read coverage 68

List of Tables

Table 1.1 Differential expression results differ greatly between versions of Cuffdiff 5
Table 1.2 Different expression results from using different reference annotations 8
Table 1.3 Increasing the read length has little effect on the results 8
Table 2.1 Overview of Cufflinks-assembled isoforms 18
Table 2.2 Classification of expressed transcripts with and without a PTC_{50nt} 18
Table 2.3 NMD-targeted genes are enriched for RNA splicing factors 23
Table 2.4 Genes with NMD-targeted isoforms are enriched for ultraconserved elements 25
Table 2.5 Features of uORF-containing transcripts that are degraded by NMD 28
Table 3.1 UPF1 mRNA depletion efficiency by experiment 40
Table 3.2 Summary of UPF1-depletion RNA-seq experiments 41
Table 3.3 Gene and isoform expression differences upon depletion of UPF1 42
Table 3.4 Effect of uORFs on transcript stability 46
Table 4.1 Summary of gene expression for the tissues 59
Table 4.2 Differential gene expression between tissues 63
Table 4.3 Alternatively spliced pharmacogenes by class 64
Table 4.4 Differential splicing between tissues 65

Acknowledgements

First and foremost, I need to thank my advisor, Steven Brenner, for giving me the opportunity to work in his lab for the past five years. With his support I was able to learn computational biology and succeed despite coming from a primarily experimental background. I was also able to gain great experiences in different fields of biology, both basic and applied, due to his collaborative nature, which I really appreciated. I had so many opportunities working in Steven's lab that I probably wouldn't have had elsewhere, including numerous collaborations, taking undergraduate intro to computer science courses, and attending and presenting at a number of major conferences. I am grateful for all of that in addition to the vital support and guidance for my research.

I'd also like to thank many members of the Brenner lab, past and present, for creating such a great and collaborative environment in which to perform research. In particular, I'd like to thank James Lloyd and Anna Desai, with whom I worked most closely with on the NMD project. Anna performed much of the experimental work that forms the basis of my analytical research and discussions with James have been an immense help with analysis and writing. Thanks also to Max Shatsky, Aashish Adhikari, and Roger Hoskins, among others, for very useful discussions. And thanks to all the Brenner lab admins for keeping the lab running. Finally, thanks to Gang Wei and Angela Brooks who were critical to my success when I was just starting in the lab - Gang for performing initial experiments and getting the NMD project started, and Angela for help and support in learning about RNA-seq analysis in particular, and about the lab and grad school life in general.

I have a number of collaborators to thank. For the NMD project, in addition to members of the Brenner lab, data was generated by Darwin Dichmann in the Harland lab and Nick Fuda in the Meyer lab at UC Berkeley, by Tom Gallagher in the Amacher lab at Ohio State, by Hiten Madhani's lab at UCSF and by Maki Inada at Cornell. Thank you to members of the Rio lab for insight and advice on analysis matters, and especially to Don himself for all the support. Thanks also to my other committee members, Britt Glaunsinger and Sandrine Dudoit, for feedback and support at committee meetings. The pharmacogenomics project was also hugely collaborative, with suggestions and insight for the analysis, in addition to the data, came from labs across the PGRN consortium including from Kathy Giacomini, Deanna Kroetz, Marisa Medina, Steve Scherer, and Wolfgang Sadee. I especially want to thank my co-authors, Aparna Chhibber and Sook Wah Yee, for all their hard work and, most of all, for making working on this project and writing the paper such a great experience. It was wonderful working with both of them.

I was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

I also want to thank friends here in Berkeley for the support and commiseration about and/or escape from grad school life. Thanks to all my MCB 2010 classmates for so many fun times, including Aisha Ellahi, Chris Alvaro, Anjali Zimmer, Jenn Cisson, Priscilla Erikson, Anne Dodson, and Nick Ellis. And thanks to my softball teammates, particularly members of the Hopeful Monsters (MCB league, Go Monsters!) and the Specials (Berkeley

league, thanks to Steph Bianco), for giving me many of my most enjoyable evenings of the past few years. I'd especially like to thank Akemi Kunibe and Michael Yormark for introducing me to different leagues and being great friends and teammates.

Finally, I need to thank my family. Thanks to my grandmother and all my aunts, uncles, and cousins, who have always believed in me. Despite living across the country, I always knew you were there for me. Thanks to Jackie, my sister-in-law, for being an awesome addition to the family. And thank you especially to my brother, Johnny, and my sister, Jacqui, for the consistent support and faith. Thank you for all the conversations, whether they were about TV shows or more serious things. Thank you for all the family vacations where we always have so much fun, no matter how much time has passed. And last, but the opposite of least, thank you to my parents, John and Nancy French. None of this would have been possible without your support throughout my life. From growing up, to going to college, to figuring out what I wanted to do next, and going to grad school, you gave me a life full of opportunities through your own hard work. Even more importantly, you gave me the belief that I could do anything because of your faith in me. I wouldn't have made it through grad school without your incredible support on every possible level. Thank you.

Chapter 1 Using RNA-seq analysis for isoform expression and alternative splicing investigations

Abstract

High-throughput RNA-sequencing (RNA-seq), the most popular of which is Illumina's short read sequencing, is now a critical method for investigating gene expression and splicing regulation. However, the technology still has limitations (primarily read length) that make analysis and interpretation of the data very difficult. Here, we survey some of the different options and tools for RNA-seq studies, focusing on isoform-level and splicing analysis. We discuss the limitations and issues in current RNA-seq analysis and offer information to help with experimental design.

Introduction to RNA-seq

High-throughput RNA sequencing, or RNA-seq, is a critical tool in investigating gene expression and alternative splicing on a genome-wide scale. Two major uses of RNA-seq are to assemble a transcriptome and to study expression changes. One can discover novel transcripts whether they are alternative isoforms or are produced by novel genes. Since the number of sequence reads is a digital readout of the number of RNA molecules in the cell, read counts can be used for differential expression and splicing analyses. However, the analysis required to do this is limited by read length and read depth.

Illumina's short read sequencing[1] is the most popular of next-generation sequencing technologies[2,3], and hundreds of millions of reads can be produced in one run, of length 50-250 bases long. While the technology has improved over the past few years from a maximum length of 100bp to 250bp reads, these reads are still very short relative to the full length of mRNA transcripts. The main disadvantage of short reads affects genes producing multiple alternative isoforms. For a given gene, most of the reads will align to sequences found in more than one isoform and this makes it difficult to determine which isoform they came from. Additionally, it is incredibly difficult to link different alternative splice events at different ends of the transcripts. In the near future, long read sequencing technologies[4-6], which can produce reads that are tens or hundreds of kilobases in length, may be a solution to these issues. Currently, however, they are currently limited by high error rate and low throughput.

In addition to read length and depth, there are a number of other options when it comes to preparing RNA-seq libraries. The primary purpose of most RNA-seq studies is to investigate mRNAs, although protocols exist for studies on smaller RNA species such as miRNAs. Library prep for mRNA-seq begins with a step to enrich for mRNAs, either by selecting for RNA molecules with poly-A tails or by depleting the samples of ribosomal RNA. Another step in library preparation is fragmentation of the RNAs and size selection. The specifics of this step primarily matter when doing paired-end sequencing, wherein the fragments (or 'inserts') are sequenced from both ends. Paired-end sequencing can mitigate some of the limits of shorter read lengths when performed on libraries with appropriately sized fragment distributions. Additionally, library preparation can be done in a strand-

specific manner so that the reads can be assigned to a particular strand, removing complications from anti-sense or overlapping transcripts.

Read alignment

Analysis of RNA-seq data, whether the goal is to find novel transcripts or to study differential expression and splicing, starts with some sort of read alignment step. This most often involves determining where a read aligns to the genome, although tools such as kallisto[7] use a pseudo-alignment step combined with transcript quantification and there are methods for analyzing RNA-seq data without the benefit of a reference genome[8-11]. There are many different tools for aligning short reads to the genome, but RNA-seq data has an added complication because reads can go across splice junctions. A number of alignment tools can handle spliced reads, either with or without the help of a reference annotation of known transcripts, including Tophat2[12,13], STAR[14], and GSNAP[15], among others[16,17]. These tools have different methods for mapping reads and for discovering novel splice junctions, and their performances in such matters vary.

A recent report evaluated of a number of alignment tools[18] and revealed that mapping percentages and the number of multi-mapped reads varies between tools and depends on such things as trimming, mismatch allowance, and how the tool deals with paired end reads. Additionally, tools that use a reference annotation as a guide do best for aligning reads to known splice sites, but the discovery of novel splice sites from just the reads is a more difficult problem and most of the tools report a large number of false positives. This can be mitigated in part by post-alignment filtering of spliced reads based on the read counts of the spliced junctions[18]. It was also reported that transcript assembly by the tool Cufflinks[19] is not terribly affected by the false positive spliced reads (see next section for more discussion). Based on some of our own work to compare Tophat2 and STAR (two of the better and faster tools), the major differences are speed (STAR is much, much faster) and STAR's capability to trim reads, which results in a higher mapping percentage for longer reads (e.g., 250bp). Ultimately though, choice of tool has little effect on downstream analysis with Cufflinks, beyond the benefit of potentially increasing mapped read depth for long reads.

Assembly of novel transcripts

As mentioned, one goal of RNA-seq analysis is to discover and assemble novel transcript isoforms. While there are a number of tools for this[19-23], it turns out to be very difficult to do with short reads, and even the best tools (Cufflinks and StringTie) do not perform well because of the complexity of extensive alternative splicing, according to a recent review[24]. The review evaluated various transcript assembly tools and found that each resulted in a substantial amount of false positives in assembling novel transcripts when using biased, imperfect (i.e., realistic) data. In particular, Cufflinks deals with the lack of full-length isoform information by reporting all combinations of alternative splicing events that are consistent with the reads. False positive assembled transcripts decrease the accuracy of downstream transcript quantification steps.

In addition to the issues with matching up different alternative events across the length of a transcript, we found that Cufflinks will report novel transcripts that use low-

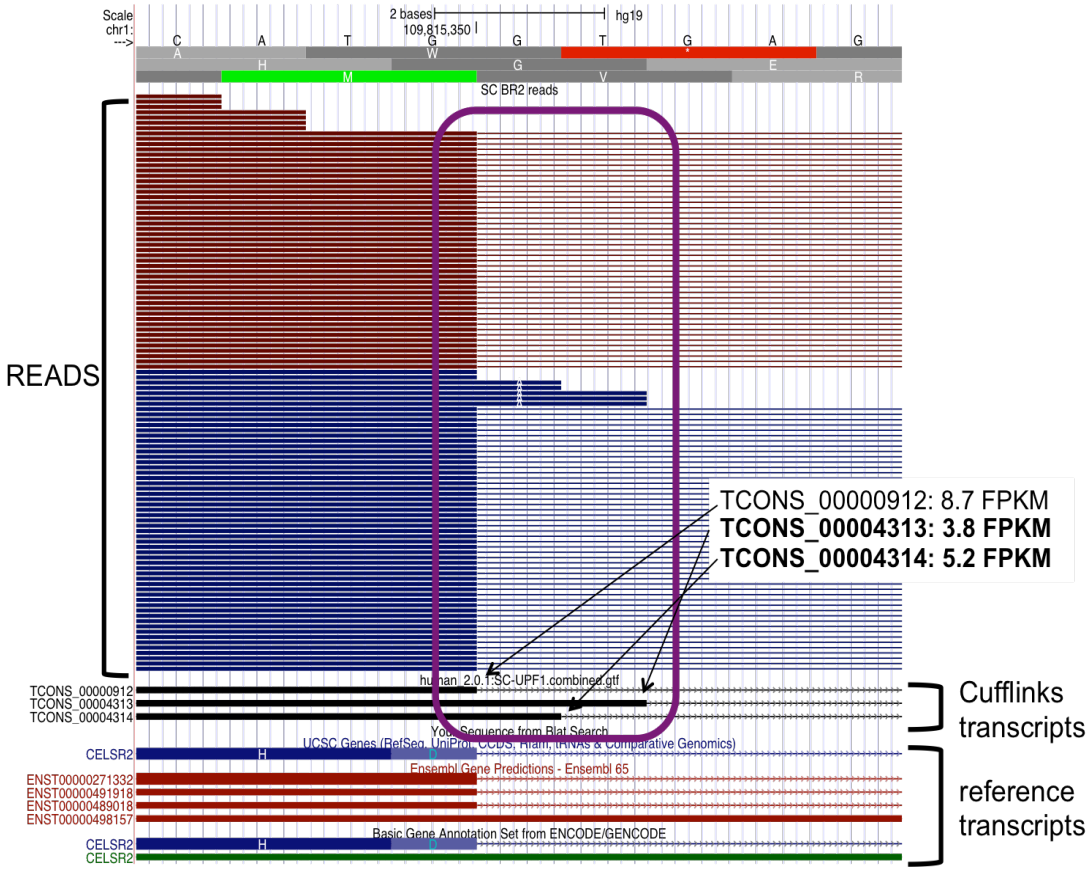
confidence junctions. Specifically, some reported novel transcripts contain splice junctions that are not supported by any spliced reads. Using transcripts assembled by Cufflinks (v2.2.1) from four samples from a human cell line (for a total of ~350 million paired end reads), of the 33,751 transcripts classified as 'novel' by Cuffmerge, 309 (0.9%) contain a splice junction with no splice read coverage. Often these are instances of extending an exon by a few bases, the only evidence for doing so being a few, likely mis-mapped reads that extend into the intron (Figure 1.1), consistent with the review paper's observation that Cufflinks has trouble sorting out the signal between exon and intron reads[24]. Of concern is the fact that we have observed Cufflinks assign substantial gene expression values to these probably non-existent transcripts and even call them significantly differentially expressed.

In order to include only novel junctions that are supported in our data, we calculate Shannon entropy scores for each splice junction using JuncBASE[25], a tool for investigating differential splicing at the junction level. Using an entropy score incorporates more information than just the junction read count; it also takes into account the number of different positions that a read that crosses the splice junction could start at. For example, a splice junction that is crossed by ten reads that all start at the same base has an entropy score of 0 - it is very unlikely that this would be the distribution of reads crossing a real splice junction (Figure 1.2). Using this metric, many low-confidence splice junctions are observed in Cufflinks-reported novel transcripts: 6,306 transcripts (19%) have a junction with an entropy score less 1, which is a score equivalent to four reads split across two start positions. Once these low confidence junctions are defined, the simplest next step is to merely filter out any transcripts that contain them before continuing with quantifying transcript abundance. However, doing this risks losing alternative splice events from other parts of the transcripts from the analysis. Another option, which we do, is try to 'fix' the low-confidence junctions by replacing them with a reference junction that shares a splice site.

Quantifying transcript expression

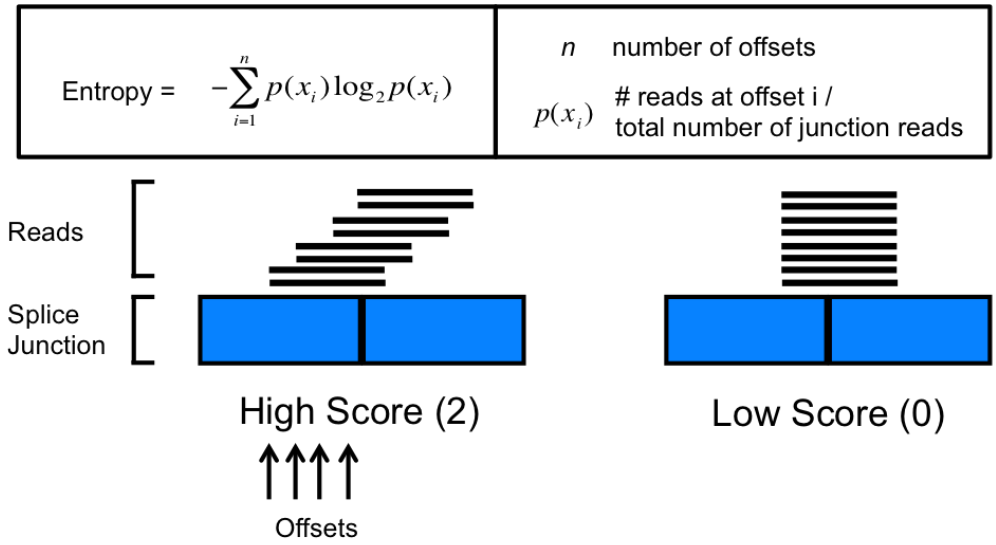
The other major use of RNA-seq data is to measure changes in gene expression and alternative splicing. There are three general methods for quantification: gene-level, isoform-level, and junction-level; each has its own uses and limitations. For studies focused solely on gene expression (as opposed to alternatively splicing), tools that perform gene-level analysis, such as edgeR[28] and DESeq[27], can be used. These involve counting the number of reads mapping to a gene and normalizing for gene length if comparing the expression to other genes, or normalizing for read depth if comparing to other conditions. The benefit of this type of analysis is that it is pretty straightforward to calculate read counts. However, potential alternative splicing of genes creates issues with calculating gene length. If a gene produces alternative isoforms of different lengths, it is unclear what should be used as the effective length of the gene, although often the union of exon lengths is used. Additionally, if substantial differential splicing occurs between conditions, this could change the effective gene length making it appear as if there is differential expression when there is not. Thus, in theory, the more accurate method of calculating gene expression is actually to first calculate expression at the transcript level and then sum the expression values of the transcripts for a gene.

Figure 1.1 Unsupported novel junctions in Cufflinks-assembled transcripts



Of the three Cufflinks-assembled transcripts (black), two have a novel junction created by the exon being extended by one or two bases. A representative subset of the reads is shown above the transcript structures (red/blue). Note the lack of junction spanning reads for the novel junctions and the mismatch (white) in the reads that do extend into the intron. While the reference transcript was assigned the highest expression value (in FPKM), the novel transcripts were also reported to have substantial expression. Cufflinks version 2.0.1[19]. Screenshot from the UCSC Genome Browser[26].

Figure 1.2 Using a Shannon entropy score to detect false positive junctions



As described in[25].

Table 1.1 Differential expression results differ greatly between versions of Cuffdiff

Cuffdiff version	# transcripts tested	# sig. increased	# sig. decreased
Version 1.0.1	156,234	28,929 (19%)	20,946 (13%)
Version 1.1.0	10,993	1,144 (10%)	942 (9%)
Version 1.2.1	42,952	17,123 (40%)	9,877 (23%)
Version 1.3.0	45,435	18,192 (40%)	10,505 (23%)
Version 2.0.1	15,998	452 (3%)	215 (1%)
Version 2.0.1*	35,573	2,000 (6%)	4,378 (12%)
Version 2.0.2	21,251	262 (1%)	98 (0.5%)
Version 2.0.2*	39,432	1,008 (3%)	2,983 (8%)
Version 2.2.1	66,752	1,136 (2%)	1,021 (2%)
Version 2.2.1*	62,549	794 (1%)	1,849 (3%)

Different versions of Cuffdiff were run on reads from an NMD-inhibition experiment (control vs experimental, two replicates). Only transcripts with a protein-coding sequence were counted for this table. A bias toward increasing in abundance is expected given the nature of the experiment (inhibition of a decay pathway). *normalized via median of geometric means as in DESeq[27], else normalization was classic FPKM.

There are a few tools that perform isoform-level quantification, including Cuffdiff[29] and RSEM[30], but there are added uncertainties in this type of analysis. The problem is that for a given gene, most exons are shared between multiple alternative isoforms and it very difficult to assign reads that map to those exons to a specific transcript. Since isoform-level quantification is difficult, and currently not very accurate[31,32] if a study is focusing on alternative splicing, junction-level quantification and differential expression may be most appropriate[25,33,34]. The benefit of this type of analysis is that it is also fairly straightforward because only non-ambiguous mapping reads can be used. However, with a focus on just one splicing event at a time, there is a loss of full-length isoform information.

There are many methods and tools for performing differential expression analyses (reviewed in[35]). The methods differ in quantification (gene vs isoform-level), over-dispersion and variability modeling, and statistical tests and corrections. Most significant genes turn out to be low expressed and less confident and so some level of expression filtering is recommended. In general, increasing the number of biological replicates increases the number of significant genes, and at least five replicates is suggested. Different versions of the same tool can produce very different results, something we have observed with Cuffdiff (Table 1.1).

Using a reference gene annotation

When using Cufflinks to discover novel alternative isoforms, we have observed that the choice of reference annotation to guide the assembly has a substantial effect on the results, particularly for quantification and differential expression, consistent with previous reports[36,37]. We investigated the use of two human gene annotations from Gencode (v19[38]) with RNA-seq data from an experiment where the nonsense-mediated mRNA decay pathway (NMD) was inhibited. Inhibition of this pathway is expected to result in an increase in abundance of transcripts with a premature termination codon (PTC). The two Gencode annotations we used were the Comprehensive Set, which includes all manual and automatic annotations (188,866 transcripts), and the Basic Set, which only includes full-length coding transcripts and excludes transcripts with a PTC (99,813 transcripts). After running the same set of mapped reads through Cufflinks and Cuffmerge (v2.2.1) with each of these two annotations, we see that the Comprehensive Set results in far more transcripts, both known and novel (Table 1.2, Row 1). However, the number of protein-coding genes is about the same (Table 1.2, Row 2), indicating the increase in number of transcripts is due to an increase in the number of alternative isoforms per gene.

For the analysis of this data and investigation of NMD, we determined the coding sequence (CDS) for each transcript. For our determination of CDS, the 'Main CDS' of a gene is the longest ORF found in any transcript from the gene. A transcript can have the 'Main CDS' (same start and stop position), a 'Main Start, Alt. Stop' CDS (same start position but resulting in a different stop codon), or a 'Alt. Start, Alt. Stop' CDS (the transcript does not overlap the start codon of the 'Main CDS' of the gene and so the longest ORF of the transcript is used). This last category is the least confident one for defining the appropriate CDS (and, arguably, has the least support for the transcript existing at all). The additional Comprehensive Set-derived transcripts fall mostly into the categories of 'Main Start, Alt. Stop' and 'Alt. Start, Alt. Stop' (Table 1.2**,** Rows 4 and 5), and so most of the increase in

number of transcripts from using the Comprehensive Set are low-confidence transcripts, whether they were in the annotation to start with, or were assembled by Cufflinks.

This increase in inclusion of low-confidence transcripts when using the Comprehensive Set results in less evidence of the expected result when we measure the biological effect of NMD inhibition. To do this, we compare the fraction of transcripts that increases >2-fold upon inhibition to the fraction that decreases >2-fold. For PTC^+ transcripts there should be a bias towards increased abundance because the degradation pathway is inhibited; PTC^- transcripts should have no biased effect since changes are likely secondary effects. Using the Comprehensive Set results in many more PTC^- transcripts, although the ratio between the fraction that increases and the fraction that decreases is the same as for the Basic Set (Table 1.2, Rows 6 and 10). On the other hand, a much larger fraction of transcripts have a PTC when using the Comprehensive Set (34% vs 18%; Table 1.2, Row 11) but a larger fraction of those decrease in abundance when NMD is inhibited which is not expected if they are true NMD targets (Table 1.2, Row 14), indicating that 1) they may not be accurately assembled isoforms, 2) the predicted CDS may be incorrect, or 3) the increase in splicing complexity for the gene decreases the accuracy of the transcript quantification. The ultimate result in using the larger set of annotations is a different picture of the outcome of NMD inhibition (Table 1.2, Row 15), and an increase in false positives for observed NMD-targeted transcripts.

Impact of read depth

When planning an RNA-seq experiment, there are numerous options for read length, paired-end vs single-end sequencing, and read depth, and often these choices result in a trade-off due to cost limitations. We have investigated the effects of these different options on gene and isoform expression analysis using the Cufflinks tools. Theoretically, using paired-end reads, longer reads, and longer fragments (with a narrower length distribution), should result in more accurate expression calculations, particularly at the isoform level. This is due to the resulting increased ability to capture more of the isoform, and therefore more splice junctions, with each fragment. However, while we do see this increase in information on the splice junction level, there is actually very little effect in the transcript expression levels reported by Cufflinks.

For example, we investigated read length by trimming 250bp paired-end reads we have from an NMD experiment to 100bp and to 50bp. Using the longer reads results in slightly more Cufflinks-assembled novel transcripts (Table 1.3, Row 2), but most of these contain low-confidence splice junctions. Thus, the overall number of confident transcripts varies very little whether the reads are 50bp long or 250bp long (Table 1.3, Row 4). The difference in length also has little effect on transcript quantification with Cuffdiff: the same fraction of transcripts are tested and increase or decrease >2-fold (Table 1.3, Rows 6-8). These negligible differences in the Cufflinks results are seen for each investigation we have performed for paired-end vs single-end, fragment size, and read length.

The other variable that we investigated for RNA-seq analysis was sequencing depth (i.e., number of fragments sequenced), which is very important for accurately quantifying transcript abundance. Again using Cuffdiff on some of our NMD data, we see that increasing

Table 1.2 Different expression results from using different reference annotations

Row	Number of transcripts	Basic Set (99,813)	Comprehensive Set (188,866)
1	Total (% novel)	143,768 (76%)	258,844 (78%)
2	CDS^+ (# genes)	94,002 (19,876)	156,023 (20,339)
3	Main CDS (% of CDS^+)	51,017 (54%)	56,262 (36%)
4	Main Start, Alt. Stop (% of CDS^+)	27,633 (30%)	56,709 (36%)
5	Alt. Start, Alt. Stop (% of CDS^+)	15,353 (16%)	43,052 (28%)
6	PTC^- (% of CDS^+)	77,286 (82%)	102,264 (66%)
7	OK to test (% of PTC^-)	17,847 (23%)	19,064 (19%)
8	>2-fold increase (% of OK)	2,980 (17%)	3,458 (18%)
9	>2-fold decrease (% of OK)	3,462 (19%)	3,999 (21%)
10	Up/Down	0.86	0.86
11	PTC^+ (% of CDS^+)	16,670 (18%)	53,717 (34%)
12	OK to test (% of PTC^+)	2,567 (15%)	5,131 (10%)
13	>2-fold increase (% of OK)	1,169 (46%)	2,157 (42%)
14	>2-fold decrease (% of OK)	334 (13%)	1,057 (21%)
15	Up/Down	3.50	2.04

Table 1.3 Increasing the read length has little effect on the results

Row		50bp reads	100bp reads	250bp reads
1	Known transcripts	204,408	204,184	204,286
2	Novel transcripts	112,485	114,163	116,392
3	Novel junctions with entropy score <1 (% of novel)	449 (0.1%)	942 (0.3%)	1,611 (0.5%)
4	Confident novel transcripts	111,317	112,223	113,033
5	Total confident transcripts	315,425	316,407	317,319
6	OK to test (% of confident)	57,537 (18%)	56,889 (18%)	58,327 (18%)
7	>2-fold increase (% of OK)	19,141 (33%)	18,961 (33%)	19,150 (33%)
8	>2-fold decrease (% of OK)	18,337 (32%)	17,400 (31%)	17,486 (30%)

the number of fragments increases the number of transcripts with enough coverage to be 'OK to test', as expected since increasing depth allows for the capture of lower expressed transcripts (Figure 1.3). Interestingly, fragment depth not only affects the number of observable transcripts, but also the number of transcripts that change in abundance. For our NMD data, we expect a bias towards transcripts that increase in abundance compared to those that decrease, since we are inhibiting a decay pathway. This expected bias disappears with lower fragment depths and requires 20-30 million fragments (40-60 million paired end reads) to be observed (Figure 1.3).

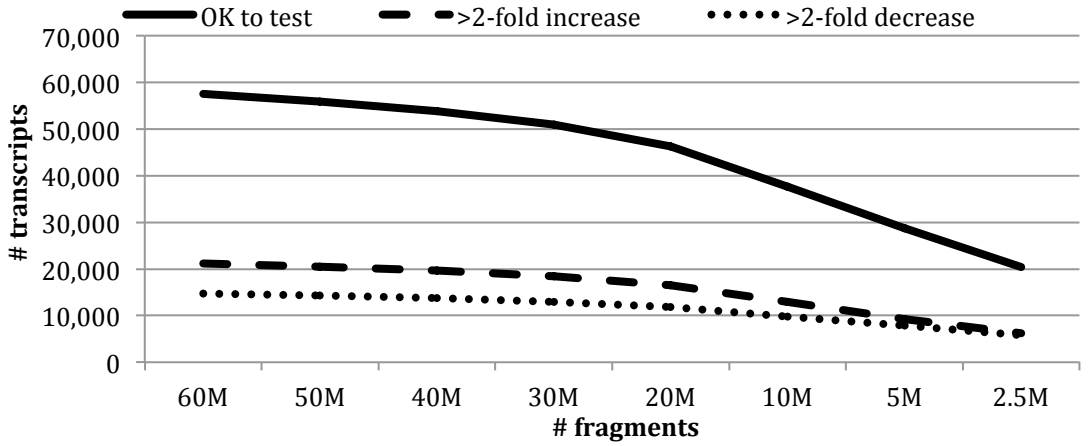
For more evidence of the importance of sufficient read depth, we investigated how it affects the convergence of transcript expression values (reported by Cuffdiff) for subsampled sets of reads from the same library. While gene expression values (calculated by summing transcript expression values) are pretty consistent between different subsampled sets even at very low fragments depths, transcript-level quantification variability reveals much more uncertainty (Figure 1.4). A depth of 20 million fragments (40 million paired end reads) is necessary for the correlation of transcript expression values calculated from subsets of reads from the same library to be above 0.8 (Spearman's rho = 0.85). These results also reveal that transcript expression quantification is still difficult even at high fragment depths: with 80 million fragments (160 million paired end reads), the correlation is still only 0.93.

Read depth versus number of samples

We also investigated the impacts of fragment depth and sample number on discovering alternative splicing events in the human population using RNA-seq data from lymphoblastoid cell lines derived from 45 individuals[39]. The 100bp paired end reads were subsampled down to 10, 20, or 40 million reads per sample and the read coverage per splicing event was calculated using JuncBASE[25]. We counted the number of splicing events with coverage of more than 5 reads/100bp, summing the coverage across samples (Figure 1.5). For example, in one sample, 7,225 splice events pass the threshold with 10 million reads, and the number increases as expected with increasing read depth: 9,665 with 20 million reads and 12,338 with 40 million reads. This pattern continues as we consider multiple samples. For 10 million reads, 13,240 events pass the threshold when the coverage is summed across different 10 samples, as do 16,551 and 19,950 events with 20 million and 40 million reads, respectively.

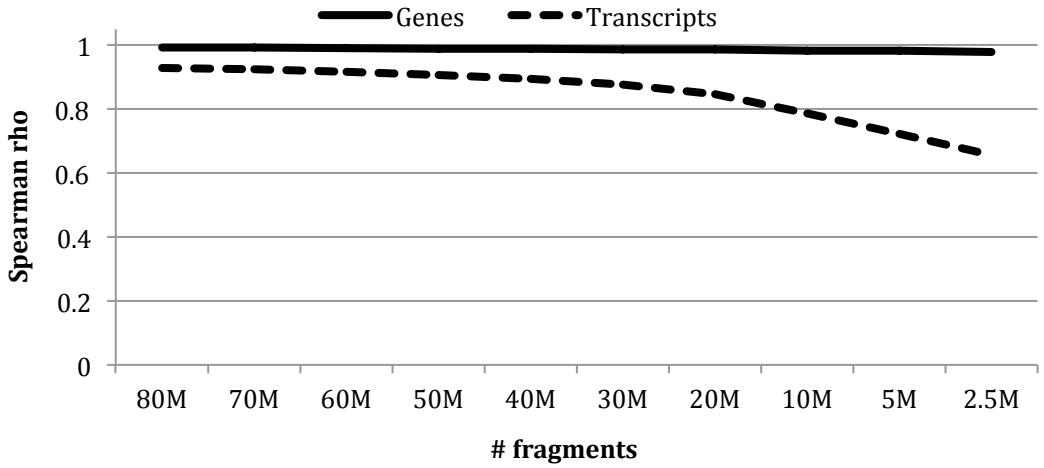
What is interesting, however, is that if the total number of reads is held constant by varying read depth and number of samples, increasing the read depth of a fewer number of samples results in the observation of more splice events than the reverse. For example, with 40 million total reads, one can sequence four samples at 10 million reads each, two samples at 20 million reads each, or one sample at 40 million reads. The latter results in 12,338 splice events, while, despite the same number of total reads, the other options result in less than 12,000 events. This difference increases as the number of total reads increases: 28 samples at 10 million reads, 14 samples at 20 million reads, and 7 samples at 40 million reads (280 million reads total) results in 15,500 events, 17,431 events, and 19,071 events respectively. Extrapolating these results, we see that there is a point where the number of extra samples able to be sequenced with the lower depth does result in more

Figure 1.3 Transcript quantification depends on fragment depth



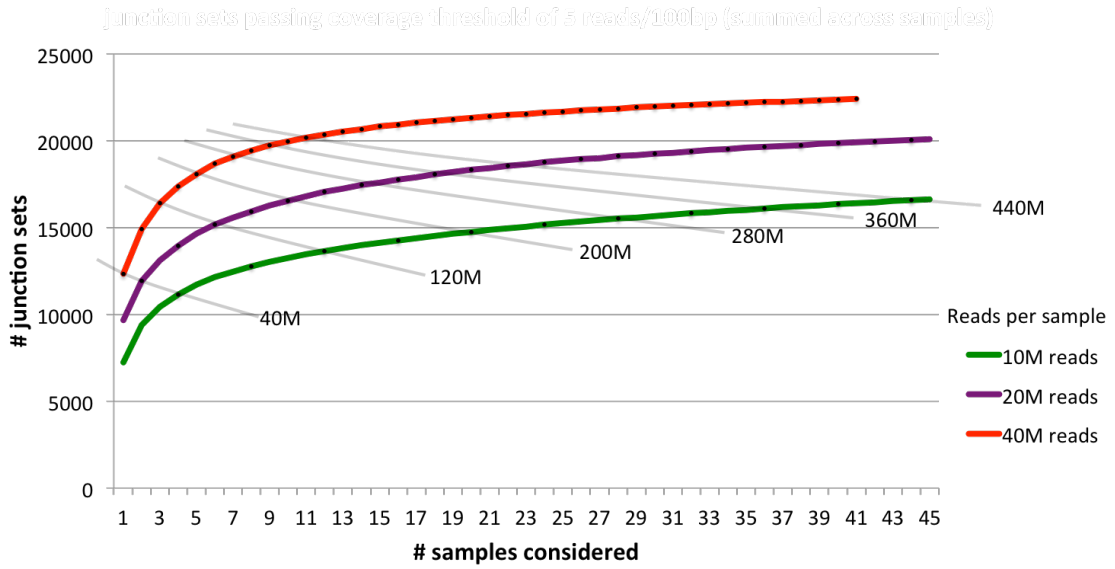
Cuffdiff (v2.2.1[29]) was run on sets of reads subsampled from 60 million fragments (120 million reads) to 2.5 million fragments (5 million reads). Solid line: the number of transcripts labeled 'OK'. Heavy dashed line: the number of 'OK' transcripts that increased >2-fold upon NMD inhibition. Light dashed line: the number of 'OK' transcripts that decreased >2-fold upon NMD inhibition. 100bp paired end reads, one replicate.

Figure 1.4 Increasing fragment depth decreases transcript quantification uncertainty



Cuffdiff (v2.2.1[29]) was run on two sets of reads each subsampled from 80 million fragments (160 million reads) to 2.5 million fragments (5 million reads). The two sets for each depth came from the same original set of reads and are mutually exclusive. The original set was generated by merging together four samples from a human cell line and is a mix of 80bp and 100bp paired end reads. Only transcripts that had an expression value >0 in at least one of the two sets (~100,000 transcripts) were used for the calculation of Spearman's *rho*.

Figure 1.5 The effect of number of samples and read depth on observing splicing events



The number of splice events observed in LCLs with increasing read depth and number of samples used. The x-axis is the number of samples considered (subsampled down from the total number of samples for the tissue). Each point is the number of splice events observed with read coverage greater than 5 reads/100bp (summed across each of the samples used), averaged after 100 permutations of subsampling.

splice events. For these data, this point is around 3 billion total reads (150 samples at 20 million reads each would reveal more splicing than only 75 samples at 40 million reads each). Similar patterns were also observed when looking at RNA-seq data from physiological tissues, such as liver, from multiple individuals[39]. Thus, if given a limited total number of reads with a goal of observing as many splice events as possible, one should generally increase read depth before increasing the number of samples. This captures more low expressed splicing events at the cost of capturing variation in splicing between individuals.

Conclusions

RNA-seq analysis is a very powerful tool for investigating gene expression and splicing regulation, however, there exist limitations in the data and the current analysis tools which require care to be taken in interpreting results[40]. While the ability to investigate the transcriptome on an isoform-level is intriguing, the difficulty of working with short reads results in uncertainty and inaccuracy in transcript assembly and quantification for even the best tools currently available[24,32]. Here we survey some of these tools and report issues we have encountered, particularly when using Cufflinks for isoform-

level investigations. We note the importance of filtering out Cufflinks-assembled transcripts that contain novel junctions that are unsupported by any spliced reads in the data, and how the choice of reference annotation and the fragment depth have a substantial effect on the quantification results. On the other hand, we saw little improvement in the Cufflinks assembly and quantification by increasing read length or using paired end reads. Finally, we observed that for junction-level analysis of alternative splicing, deeper read depth of a smaller set of human samples results in more observable junctions than sequencing more samples at shallower depth.

References

1. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
2. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
3. Bentley, D. R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16,** 545–552 (2006).
4. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4,** 265–270 (2009).
5. Pennisi, E. Genomics. Semiconductors inspire new sequencing technologies. *Science (New York, N.Y.)* **327,** 1190 (2010).
6. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5,** 829–834 (2008).
7. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3519
8. Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30,** 1660–1666 (2014).
9. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28,** 1086–1092 (2012).
10. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).
11. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8,** 1494–1512 (2013).
12. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
13. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
15. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).
16. Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F. & Rätsch, G. RNA-Seq read

alignments with PALMapper. *Curr Protoc Bioinformatics* **Chapter 11,** Unit 11.6 (2010).
17. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38,** e178 (2010).
18. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–1191 (2013).
19. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
20. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).
21. Li, W., Feng, J. & Jiang, T. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* **18,** 1693–1707 (2011).
22. Bernard, E., Jacob, L., Mairal, J. & Vert, J.-P. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics* **30,** 2447–2455 (2014).
23. Mezlini, A. M. *et al.* iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* **23,** 519–529 (2013).
24. Hayer, K. E., Pizarro, A., Lahens, N. F., Hogenesch, J. B. & Grant, G. R. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics* **31,** 3938–3945 (2015).
25. Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.* **21,** 193–202 (2011).
26. Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39,** D876–82 (2011).
27. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
28. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
29. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7,** 562–578 (2012).
30. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
31. Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16,** 150 (2015).
32. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17,** 74 (2016).
33. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
34. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).
35. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinformatics* **16,** 59–70

(2015).
36. Zhao, S. & Zhang, B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* **16,** 97 (2015).
37. Leshkowitz, D. *et al.* Using Synthetic Mouse Spike-In Transcripts to Evaluate RNA-Seq Analysis Tools. *PLoS ONE* **11,** e0153782 (2016).
38. Coffey, A. J. *et al.* The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.* **19,** 827–831 (2011).
39. Chhibber, A. *et al.* Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *Pharmacogenomics J.* (2016). doi:10.1038/tpj.2015.93
40. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17,** 13 (2016).

Chapter 2 Transcriptome Analysis of Alternative Splicing Coupled with Nonsense Mediated mRNA Decay in Human Cells

Abstract

To explore the regulatory potential of nonsense-mediated mRNA decay (NMD) in human cells, we globally surveyed the transcripts targeted by this pathway via RNA-Seq analysis of HeLa cells in which NMD had been inhibited. We first identified those transcripts with both a premature termination codon more than 50 nucleotides upstream of an exon-exon junction (50nt rule) and a significant increase in abundance upon NMD inhibition. Remarkably, at least 2,793 transcripts derived from 2,116 genes are physiological NMD targets (9.2% of expressed transcripts and >20% of alternatively spliced genes). Our analysis identifies previously inferred unproductive isoforms and numerous previously uncharacterized ones. NMD-targeted transcripts were derived from genes involved in many functional categories, and are particularly enriched for RNA splicing genes as well as ultraconserved elements. By investigating the features of all transcripts impacted by NMD, we find that the 50nt rule is a strong predictor of NMD degradation while 3' UTR length generally has only a small effect in human cells. Additionally, thousands more transcripts without a premature termination codon in the main coding sequence contain a uORF and display significantly increased abundance upon NMD inhibition indicating potentially widespread regulation through decay coupled with uORF translation. Our results support that alternative splicing coupled with NMD is a prevalent post-transcriptional mechanism in human cells with broad potential for biological regulation.

Introduction

Nonsense-mediated mRNA decay (NMD) is a eukaryotic mRNA surveillance pathway with a role in gene regulation. In its surveillance capacity, NMD recognizes premature termination codons (PTCs) generated by nonsense mutations or errors in transcription and splicing and eliminates the aberrant transcripts in order to protect the cell from the production of potentially harmful truncated proteins[1-5]. Additionally, many transcripts targeted to NMD are generated by highly regulated and evolutionarily conserved splicing events[6-10]. This coupling of alternative splicing and NMD (AS-NMD) can be used to post-transcriptionally regulate the level of mRNA (also termed Regulated Unproductive Splicing and Translation, or RUST)[11-15] .

A number of splicing factors are known to auto-regulate their own expression or regulate the expression of other splicing factors through alternative splicing coupled with NMD[16-25]. Many of these NMD-coupled splicing events (including all those in the SR gene family) are associated with a highly conserved or ultraconserved element of the human genome, indicating a potential role for these elements in the regulation of alternative splicing coupled with NMD[7,26]. Genes with other functions have also been reported to be regulated via alternative splicing coupled to NMD such as the spermidine/spermine N1-acetyltransferase (SSAT) gene which has been shown to produce less of the NMD-targeted isoform when the protein is needed in the presence of its substrate[27,28].

Numerous studies have been performed to investigate the components modulating the NMD process, resulting in identification of many NMD factors[29-38]. Among those proteins known to mediate the NMD pathway, UPF1 is a crucial and conserved component[39-44]. However, the specific mechanism(s) of NMD and the defining features of NMD targets are still not completely understood and likely vary across different species[45]. In mammals, the presence of an exon-exon junction more than 50 nucleotides downstream of a stop codon induces NMD (the 50nt Rule)[46,47]. The exon junction complex (EJC) is a protein complex deposited near an exon-exon junction during the splicing process and acts as a cellular memory of an intron's location. When a ribosome terminates upstream, the EJC recruits NMD factors including UPF1 and triggers decay. Long 3' UTRs have also been found to target transcripts to NMD[1,48-53]. A long 3' UTR has been proposed to trigger NMD in a couple of ways: structural characteristics, such as the long distance between PABP on the poly-A tail and the terminating ribosome, lead to aberrant termination[1,54] or that UPF1 binds to 3' UTRs in a length-dependent manner and this primes the transcripts for NMD[55].

The potential prevalence of NMD-targeted alternative isoforms was first established by searching EST databases for transcripts with a PTC[11]. These datasets, however, were generated from cells where NMD was active and the targets should be depleted. There have also been studies that successfully identified NMD-targeted transcripts by inhibiting NMD (often by perturbing UPF1) and using microarray methods[25,50,56-59]. However, microarray analyses generally rely on probes derived from known transcript sequences, again, derived from normal cells where NMD is active. Thus, their capacity to detect native transcripts targeted by NMD is still limited, and many novel isoforms were likely missed. High-throughput deep sequencing of the transcriptome and advanced analytical algorithms now allow for the identification and quantification of novel isoforms. These methods have been combined with NMD inhibition and used to investigate the effect of NMD on the transcriptome in mouse[52,60] and in human[38,61] and deep sequencing of mouse and human brains has revealed extensive alternative splicing coupled to NMD, even in the presence of active NMD[9].

Many questions regarding the prevalence and mechanism of NMD remain open. In this study we used NMD inhibition via knockdown of UPF1 and RNA-seq analysis in a human cell line (HeLa) in order to discover genes that produce putative unproductive isoforms and investigate the features of transcripts affected by NMD. For both known and novel transcripts, we predicted isoform-specific putative PTCs (by the 50nt Rule), measured isoform-level expression changes, and found that a substantial fraction of alternatively spliced transcripts are subject to NMD, even in just this single cell type and condition. Furthermore we were able to investigate the features of NMD-targeted isoforms and found that while the presence of an exon-exon junction ≥50nt downstream of a stop codon was a strong predictor of NMD susceptibility, the length of the 3' UTR on its own had a much smaller effect globally. As seen before, we also found that NMD targets are enriched for the presence of ultraconserved elements and for splicing regulators and RNA-binding proteins. Additionally, many transcripts with uORFs are sensitive to NMD. Our findings include numerous previously unreported substrates for this process in human cells, implying broad regulatory potential.

Results

Over two thousand genes produce transcripts identified as confident NMD targets

Transcripts with a premature termination codon (PTC) should stabilize if the NMD pathway is inhibited. In order to inhibit NMD in HeLa cells, we used short hairpin RNAs (shRNAs) against the mRNA of NMD factor UPF1[62,63]. For each biological replicate, a paired control experiment using an shRNA of random sequence was also performed[62]. Western blot analysis confirmed that the protein level of UPF1 in the knockdown cells was about 6% of that in the control cells. Furthermore, using real-time PCR, we measured the abundance of known NMD-targeted isoforms of two SR genes (*SRSF2* and *SRSF6*), and both showed a 4- to 30-fold increase in mRNA abundance in the UPF1-depleted cells, suggesting that NMD was substantially inhibited.*

To systematically survey NMD-targeted transcripts, we constructed strand-specific RNA-Seq libraries from mRNA of cells with inhibited NMD and mRNA of control cells for two biological replicates of the UPF1-knockdown experiment. Over 40 million 80bp paired-end reads were obtained for each sample of one biological replicate by sequencing with an Illumina GA IIx, and 180 million 101bp paired-end reads were obtained for each sample of the other replicate on a HiSeq 2000. Tophat[64] was used to map over 80% of reads to the human genome. We used Cufflinks[65,66] to assemble the reads into known and novel isoforms for each sample and merged these sets to yield a universal set of 131,003 isoforms. We used the Cufflinks sub-tool Cuffdiff to quantify transcript abundance and determine significant differential expression between inhibited NMD cells and control cells. We filtered out isoforms expressed at a low level (FPKM<1) in both conditions in order to be more confident in the quantification, resulting in data for 60,976 substantially expressed isoforms (Table 2.1).

To determine which isoforms have a putative PTC, we predicted the coding sequence (CDS) of each isoform for each gene. Ultimately, 30,317 expressed isoforms derived from 11,055 genes have a predicted CDS according to our method (Table 2.1). We first sought a list of high-confidence direct NMD targets, and so we focused on those transcripts with a termination codon more than 50 nucleotides upstream of the transcript's last exon-exon junction - a putative PTC according to the 50nt rule (PTC_{50nt}) that is expected to trigger NMD. We found that a fifth (6,429) of the CDS-containing transcripts contained a PTC_{50nt}.

Direct targets of NMD should also exhibit significantly higher abundance when NMD was inhibited. In our present study, we identified 10,717 CDS-containing transcripts that increased at least 1.5-fold in the inhibited NMD sample and had significant differential expression according to Cuffdiff (Table 2.2). These were significantly enriched for PTC_{50nt}-containing transcripts (Fisher's exact test, $p < 2 \times 10^{-308}$). Altogether, 3,832 out of 6,429 (59.6%) PTC_{50nt}-containing transcripts were significantly more abundant in NMD inhibited cells, compared to only 871 (13.5%) that were significantly less abundant (Table 2.2, Figure 2.1A). Many other PTC_{50nt}-containing transcripts demonstrated increased

* This experimental work was performed by Gang Wei while in Steven Brenner's lab.

Table 2.1 Overview of Cufflinks-assembled isoforms

Class Description	# of isoforms	Expressed isoforms (FPKM >1 in at least 1 sample)	CDS-containing isoforms
Matches reference exactly (=)	74,233	23,687	21,616
Contained with reference (c)	1,627	628	391
Novel isoform of reference (j)	16,993	11,413	8,160
Unknown, intergenic (u)	16,232	10,515	0
Probably pre-mRNA (e)	575	279	72
Completely intronic (i)	13,993	9,422	0
Other generic exon overlap (o)	467	249	73
Apparent polymerase run-on (p)	3,156	1,824	0
Exon overlap, antisense (x)	773	393	0
Intronic, antisense (s)	0	0	0
Repeat (r)	0	0	0
Multiple classes (.)	2,954	2,565	5
TOTAL	131,003	60,976	30,317

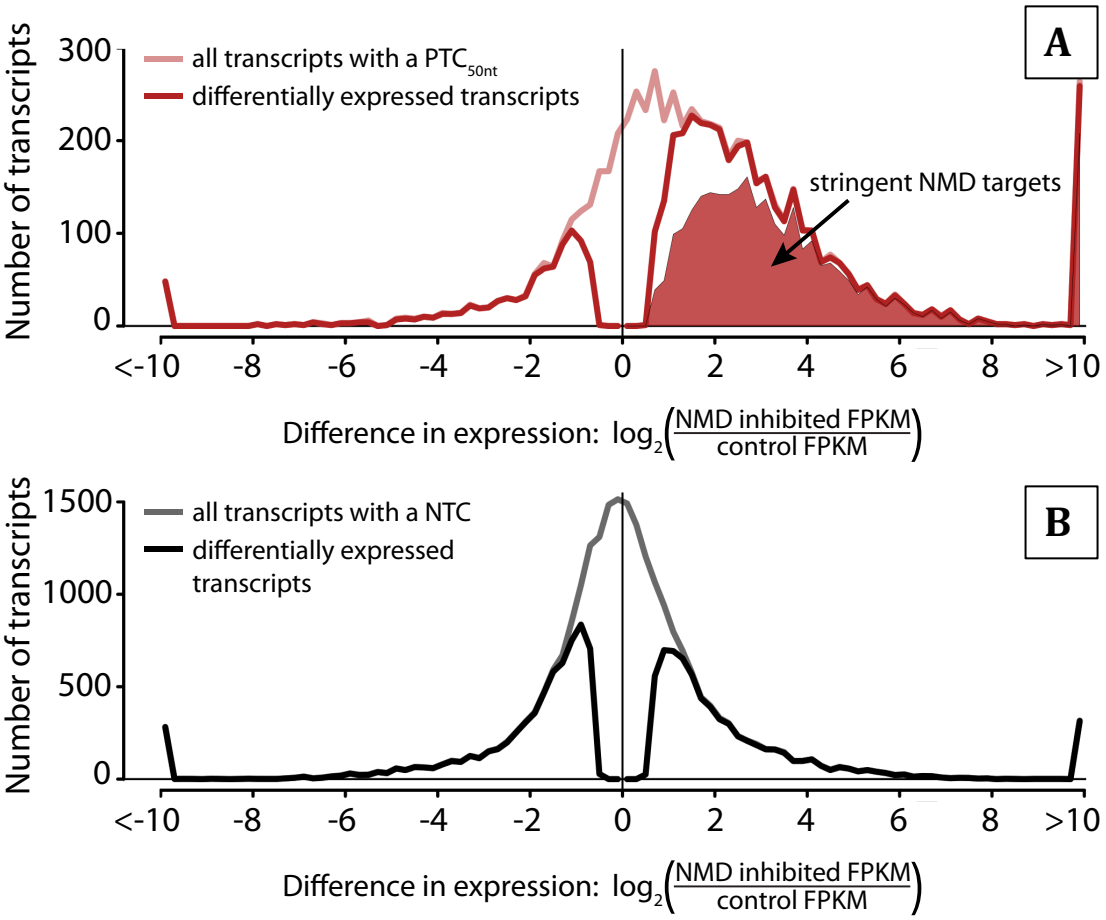
Note: the code in parentheses is used by Cufflinks (v1.0.1) to classify assembled transcript.

Table 2.2 Classification of expressed transcripts with and without a PTC_{50nt}

	Number of isoforms	Increased abundance	Decreased abundance
All expressed transcripts with a defined coding sequence	30,317	10,717 (35%)	7,600 (25%)
Transcripts with a NTC	23,888	6,885 (29%)	6,729 (28%)
Transcripts with a PTC_{50nt}	6,429	3,832 (60%)	871 (14%)

Note: Increased and decreased abundance refers to the expression changes in NMD-inhibited cells (compared to control cells) that are significant according to Cuffdiff and >1.5 fold. The percentage is of the number of transcripts in that category.

Figure 2.1 Transcripts with a PTC_{50nt} are more likely to increase when NMD is inhibited

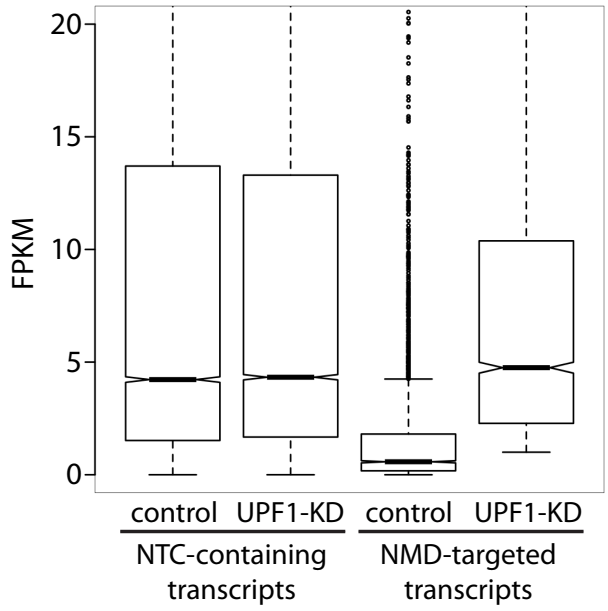


A. Histogram of expression changes of PTC_{50nt}-containing transcripts between NMD inhibited cells and control cells. B. Histogram of expression changes of NTC-containing transcripts between NMD inhibited cells and control cells. Transcripts that were not expressed in one of the conditions were given a $\log_2$(fold change) value of ±10. PTC_{50nt}-containing transcripts demonstrated a strong bias towards accumulation when NMD was inhibited (Kolmogorov-Smirov (KS) test, $p < 2 \times 10^{-308}$).

abundance, but were not statistically significant. Compared to transcripts with a non-PTC_{50nt} termination codon (an NTC), PTC_{50nt}-containing transcripts demonstrated a strong bias towards accumulation when NMD was inhibited (Kolmogorov-Smirov (KS) test, $p < 2 \times 10^{-308}$). Transcripts with an NTC had similar percentages of significantly increased (28.8%) and decreased (28.2%) transcripts (Figure 2.1B), suggesting that they may be affected by secondary effects from the knockdown of the UPF1 protein and are unlikely to have been directly affected by NMD.

Since UPF1 has multiple functions, including some independent of NMD [67-69], its depletion could cause gene expression changes at the transcriptional level. Additionally, mis-regulation of genes regulated via alternative splicing coupled to NMD could have downstream effects. To minimize the potential contribution of these transcriptional

Figure 2.2 Distribution of expression levels of NMD targets and non-NMD targets



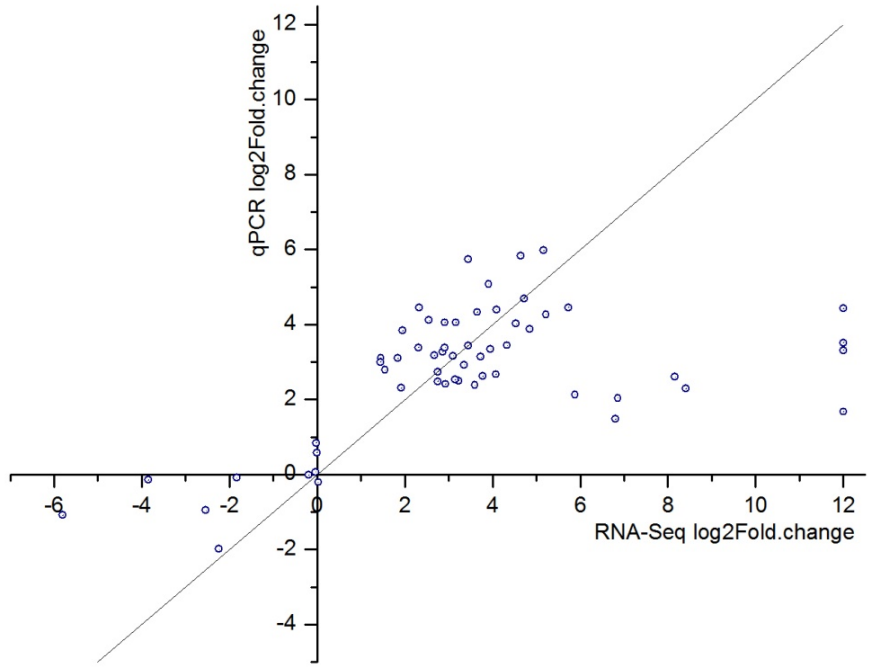
The distributions of FPKM values for transcripts without a premature termination codon (NTC-containing) and for NMD-targeted transcripts in control cells (NMD active) and UPF1-knockdown cells (NMD inhibited). The plots were generated by the R boxplot function and show the median and upper and lower quartiles.

changes to the increased abundance of PTC_{50nt}-containing transcripts, we further focused on PTC_{50nt}-containing transcripts that were significantly 1.5-fold more abundant in NMD inhibited cells (and increased in both biological replicates independently), and also adhere to either of the following criteria: 1) no NTC-containing isoform from the same gene was more than 1.2-fold higher in the inhibited NMD sample, or 2) the PTC_{50nt}-containing transcript abundance increase was at least 2-fold higher than the increase of the NTC-containing isoforms from the same gene. Note that these criteria require that the genes in question also produce an NTC-containing isoform in this experiment (FPKM>0 in at least one sample). This will exclude genes that are constitutive NMD targets such as some selenoproteins[70,71] and the NMD factors themselves[50,72].

We thus identified 2,793 isoforms (from 2,116 genes) as our high confidence set of direct NMD targets. These transcripts were derived from 19.1% of expressed, protein-coding genes, and 22.4% (2,116 out of 9,440) of alternatively spliced genes in this cell type (alternatively here defined as producing 2 or more isoforms, including the NMD targets). Before degradation, many of these NMD-targeted transcripts are produced at levels comparable to those of NTC-containing transcripts (Figure 2.2), suggesting that they are not merely splicing noise but instead play an important role in regulation. To confirm the expression changes reported by our analysis of the RNA-Seq data, 48 NMD-targeted transcripts were checked by qPCR and all had >1.5-fold increase when NMD was inhibited, consistent with our RNA-seq results (Figure 2.3*).

* This experimental work was performed by Gang Wei while in Steven Brenner's lab.

Figure 2.3 qPCR validation of NMD targets

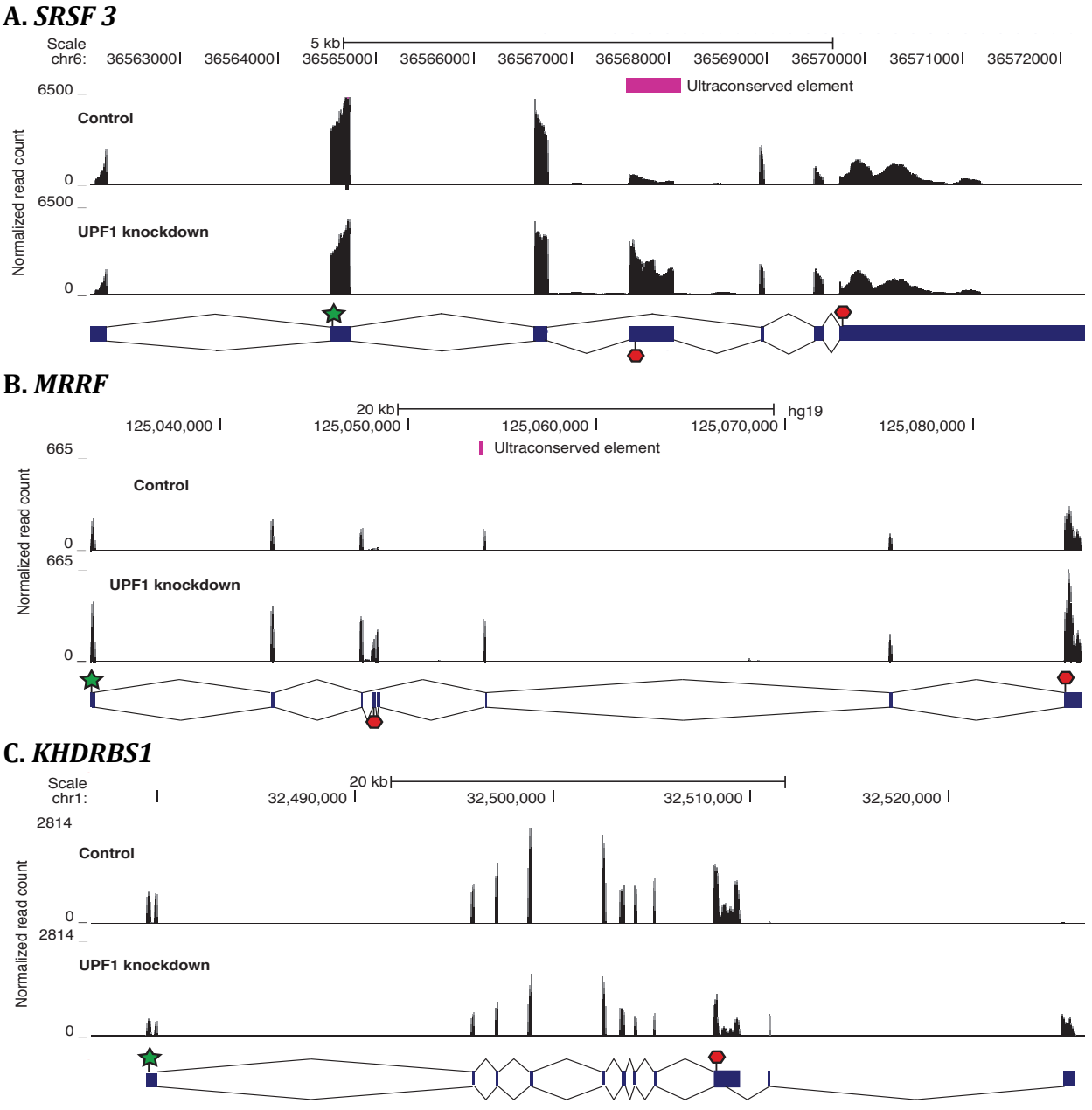


Log_2(UPF1-knockdown/control expression) of isoforms in RNA-seq data vs qPCR. Forty-eight high-confidence NMD targets (points on right side of plot) were tested and all of them increased >2-fold when measured by qPCR. Note that the four isoforms with a log_2(fold change) of 12 in the RNA-seq data were expressed below the level of detection in the control sample (FPKM=0, therefore true fold change is actually infinity but was capped at 12 in this analysis). Additionally, the five isoforms reported to increase substantially more in the RNA-seq data than by qPCR were expressed at very low levels in the control cells (all had FPKM<1, three had FPKM< 0.2) and this would make it more difficult to accurately quantify their abundance in the control samples and may explain the over-estimation of fold change in RNA-seq data. However, they all still substantially increased when NMD was inhibited using qPCR. Ten non-NMD targeted isoforms were also measured by qPCR: 5 that did not change in the RNA-seq data (clustered around 0 in plot) and 5 that decreased (left side of plot). Experiment performed by Gang Wei.

Diverse categories of genes are affected by NMD, particularly splicing

A number of human genes have been previously reported to produce NMD-targeted transcripts, including genes encoding splicing factors and NMD factors. Many of the genes with previously validated NMD-targeted isoforms were found in our results. Specifically, we found that the majority of the genes encoding SR proteins expressed in our data generated an NMD-targeted isoform (e.g., Figure 2.4A), as previously reported[7,26]. The only SR gene without a high-confidence NMD-targeted isoform in our study is *SRSF9*, which was reported as having a PTC-containing cassette exon based on homology to mouse. A closer investigation showed that there were only a few junction reads supporting the poison cassette exon in our present data, preventing the confident assembly of the isoform. We also found that five NMD factors (*UPF2*, *SMG1*, *SMG5*, *SMG6*, *SMG7*) reported to be auto-regulated[50,72] all showed significant gene expression increases of 1.6- to 3.3-fold when UPF1 was knocked down, while *UPF3A* and *UPF3B* showed non-significant gene expression

Figure 2.4 Known and novel splicing events resulting in NMD targets



Normalized read counts for data from control cells (top track) and NMD-inhibited cells (bottom track). Isoform structures are shown below the read tracks. The stop indicated underneath the exon is the PTC. A magenta box above the read tracks indicates the presence of an ultraconserved element.

change (<1.2-fold). Additionally, our data indicate that *SMG7* and *UPF3A* produce NMD-targeted alternative isoforms.

Expanding on the set of splicing-related genes producing NMD-targeted isoforms, we found that a total of 139 genes annotated with the Gene Ontology (GO) term[73] 'RNA binding' and 73 genes annotated with 'RNA splicing' produce NMD targets (Table 2.3). These categories are in fact significantly enriched in NMD-targeted genes (Fisher's exact test FDR<0.05; background is genes with an expressed isoform (FPKM>1 in at least one sample)), along with 'mitochondrion', 'methyltransferase activity', and 'metabolic process' (Figure 2.5). The enrichment of the category 'mitochondrion' is in agreement with a report that genes affected by UPF2 knockout in mouse liver cells are enriched for the 'mitochondrion' category[60]. While not significantly enriched, we found that over 4,000 different GO categories contained genes generating NMD-targeted transcripts, indicating NMD-targeted genes are involved in a number of diverse functions such as protein folding, translation, oxidation-reduction processes, vesicle-mediated transport, and chromatin modification.

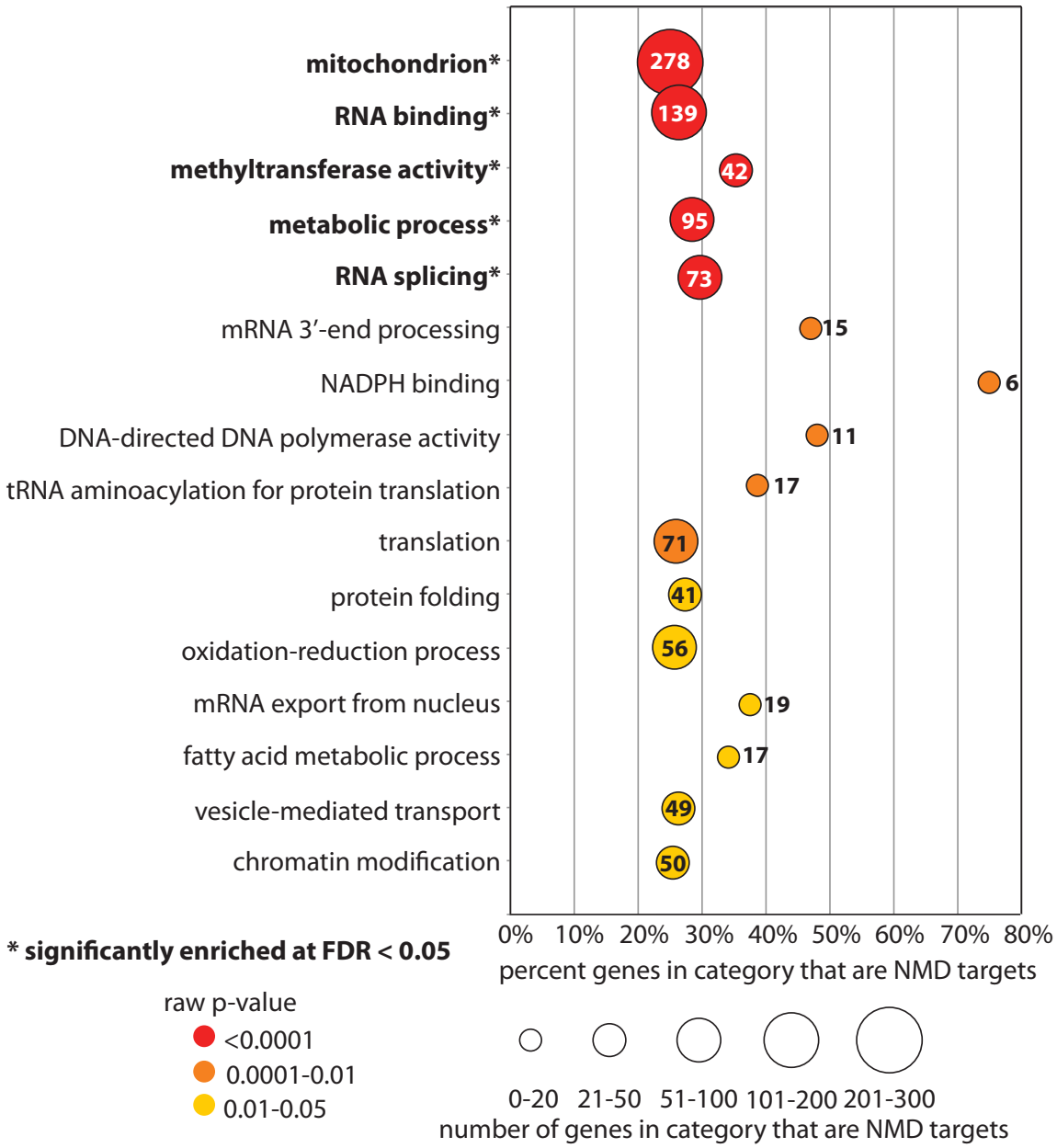
We further found that 71 translation-related genes had NMD-targeted transcripts, including many genes encoding ribosomal proteins (34). This is likely an underestimation of the number of ribosomal genes that are targeted by NMD since most ribosome genes are transcriptionally down-regulated in our NMD-inhibited cells (130 of 169 genes annotated 'ribosome' are >1.2-fold down) and this may mask the increased abundance of NMD-targeted isoforms. In total, over a third of expressed ribosomal genes (58 of 169) have an observable alternative isoform with a PTC_{50nt} and are potential NMD targets, even if they don't fall into our high-confidence set.

Table 2.3 NMD-targeted genes are enriched for RNA splicing factors

Splice factor category	Genes with isoforms degraded by NMD
SR	SRSF1, SRSF2, SRSF3, SRSF4, SRSF5, SRSF6, SRSF7, SRSF8, SRSF10, SRSF11
hnRNP	CIRBP, HNRNPA2B1, HNRPDL, HNRNPH1, HNRNPH3, HNRNPK, PTBP1, PTBP2, RBM3
snRNP	PPIH, PRPF3, SART1, SNRNP40, SNRNP48, SNRNP70, TXNL4A, U2AF1, U2AF2
DEAD	DDX5, DDX46, DHX9, DHX15, DHX35, INTS6
Sm	SNRPA1, SNRPB, SNRPN
Other	ACIN1, AKAP17A, CCAR1, C16orf80, CDK12, CIR1, CLASRP, CLK1, CPSF1, CRNKL1, CSTF1, DGCR14, DNAJC8, EIF2S2, FUBP3, FUS, GCFC1, GTF2F2, IVNS1ABP, LUC7L3, MAGOHB, MOV10, NCBP2, POLR2E, PRPF4, PRPF4B, PRPF18, PRPF38B, RBM25, SCAF8, SF3B1, SFPQ, SFSWAP, SMNDC1, SRPK1, SREK1, SRRM1, SRRM2, SUGP1, SUGP2, TCERG1, THOC2, THOC4, TIA1, TIAL1, TOP1MT, TRA2A, TRA2B, TXNL4B, U2SURP, USP39, YTHDC1, ZCCHC8, ZFR, ZNF207, ZRSR2

Genes annotated with the GO term 'RNA splicing'. Underlined genes were previously found to have NMD-targeted isoforms.

Figure 2.5 NMD-targeted genes fall into diverse functional categories



Genes with an isoform degraded by NMD were classified into Gene Ontology functional categories. The asterisk indicates the category is significantly enriched for genes with NMD-targeted isoforms (Fisher's exact test FDR<0.05; background is genes with an expressed isoform (FPKM>1 in at least one sample)).

Table 2.4 Genes with NMD-targeted isoforms are enriched for ultraconserved elements

Functional category		Genes with isoforms targeted by NMD and exonic UCE
RNA processing	16	DDX5, DHX15, HNRNPH1, HNRNPK, HNRPDL, PRPF38B, PTBP2, SRSF1, SRSF3, SRSF6, SRSF7, SRSF11, TIAL1, TRA2A, TRA2B, ZFR
Transcriptional regulation	4	CCAR1, MED1, MGA, NFAT5
Other	6	HIRA, RC3H2, FAM98A, MRRF, STRN3, DLG2

Underlined genes: UCE overlaps the alternative splicing event that generates the NMD-targeted isoform.

NMD-targeted genes are enriched for ultraconserved elements

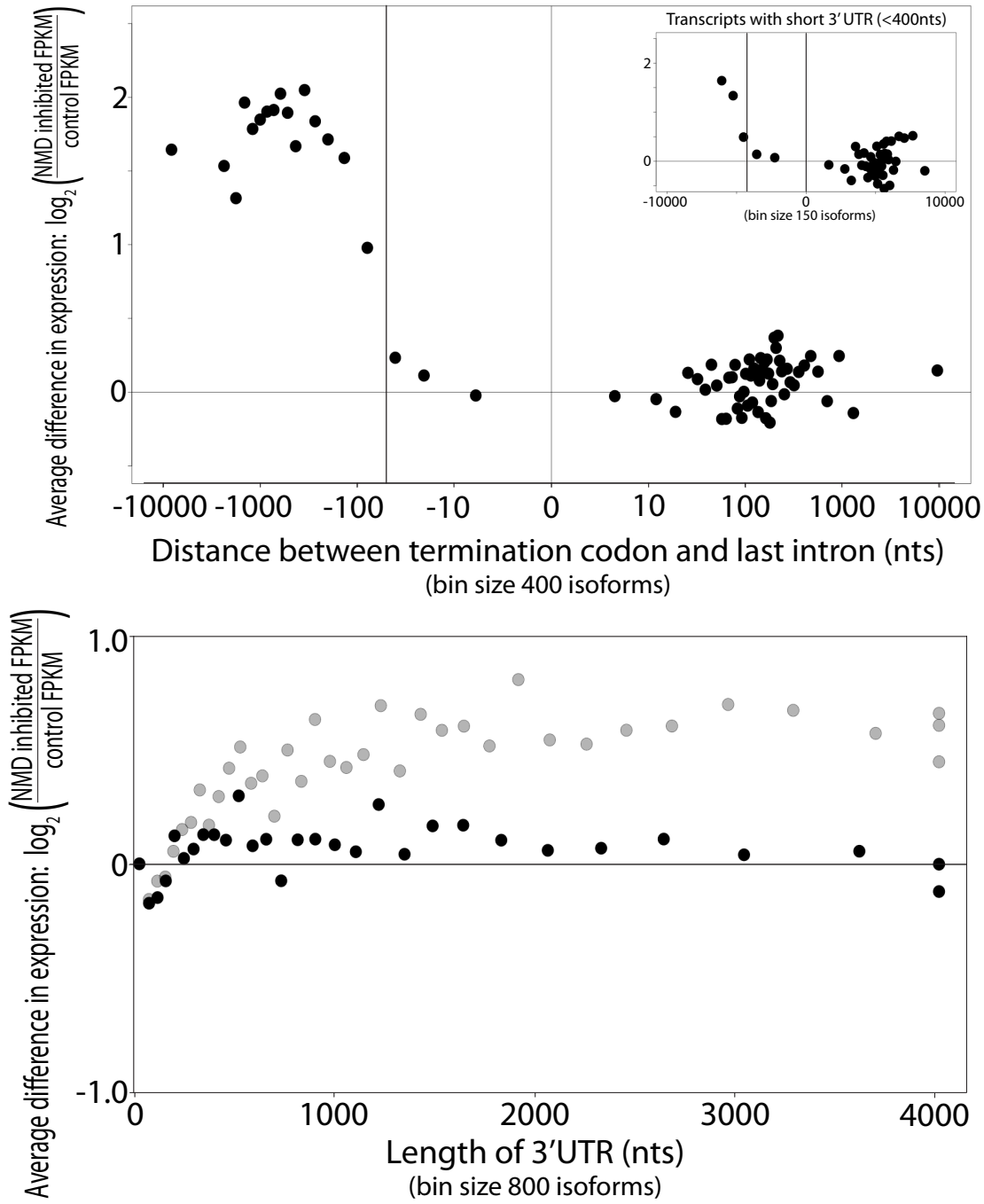
The splicing events that generate PTCs in the SR genes are associated with highly conserved nucleotide sequences[6,7], including those termed ultraconserved elements (UCEs)[74]. To determine if ultraconserved sequences are more generally enriched in genes producing NMD-targeted isoforms, we examined the overlap between 481 reported UCEs and our high-confidence set of NMD targets. We found that 175 UCEs overlapped 120 expressed genes in our data (83 of these overlapped an exon). Of the 2,116 genes producing an NMD-targeted isoform, 26 overlapped an exonic UCE and 12 genes overlapped an intronic UCE. NMD targets are significantly enriched for exonic UCEs (Fisher's exact test, $p = 6.7 \times 10^{-4}$) but not for intronic UCEs (Fisher's exact test, $p = 0.24$).

Intriguingly, for at least 21 of these genes, the UCE covers the alternatively spliced region that generates the PTC, supporting the hypothesis that UCEs maybe involved in splicing regulation. Most of these alternative splicing events are cassette exons. Interestingly, we found that the majority of the NMD-targeted genes associated with UCEs encode RNA-binding proteins involved in pre-mRNA processing, but those with other functions such as signaling and transcriptional regulation were also found (Table 2.4). One typical example is MRRF, a mitochondrial ribosome recycling factor involved in ribosome release at translation termination, whose alternative isoform has a multiple cassette exon inclusion event that generates a PTC. The downstream exon and relevant intron overlap 201nt of 100% conservation between human and rodent (Figure 2.4B). Some UCEs completely contained within an intron of an NMD-targeted gene may also play a role in regulating splicing to generate PTC-containing isoforms, although the mechanism is less clear.

The 50nt Rule is a strong predictor of NMD while a longer 3' UTR has little effect

While 60% of transcripts containing a PTC_{50nt} were significantly more abundant when NMD was inhibited (and over 70% increased at least 1.2-fold), a PTC_{50nt} is not the only trigger of NMD in human cells. The length of the 3' UTR has also been reported to have an effect on whether a transcript is degraded by NMD. To investigate the correlation between downstream exon-exon junctions ('50nt rule') and NMD, we compared the

Figure 2.6 The 50nt Rule is a strong predictor of NMD and a longer 3' UTR is not



A. The distance the termination codon is downstream of the last exon-exon junction versus $\log_2$ fold change in expression when NMD is inhibited compared to control cells. Each point is an average on both axes of 400 isoforms. Transcripts with distances greater than ±10kb fall into the first or last bins. B. Length of the 3' UTR versus $\log_2$ fold change in expression when NMD is inhibited compared to control cells. Gray: all expressed transcripts. Black: Only expressed NTC-containing transcripts. Each point is an average on both axes of 800 isoforms. Transcripts with 3' UTRs longer than 4kb fall into the last bin.

distance between the termination codon and the last exon-exon junction of the transcript with the change in abundance when UPF1 is knocked down. We found that transcripts with a PTC_{50nt} are twice as likely to increase than those without, and on average they increase more than 3-fold while on average those without a PTC_{50nt} do not change (Figures 2.6A). Even for transcripts with a short 3' UTR (<400nts), we see a strong correlation of NMD with a downstream exon-exon junction (Figure 2.6A inset). Additionally, there is a clear shift in likelihood of degradation right at a 50nt distance, as expected. Our results show a stronger effect than the results reported by Hurt, et al[52] in mouse cells. While they report only a 1.2-fold abundance increase of PTC_{50nt} transcripts compared to NTC transcripts, we see a 2.6-fold increase.

While we see some correlation between increased 3' UTR length and increased abundance, this is mostly explained by a strong correlation between 3' UTR length and likelihood of a 3' UTR intron and largely disappears when looking at only NTC-containing transcripts (Figure 2.6B). Without downstream exon junctions, transcripts with a 3' UTR longer than 2000nts are significantly more likely to increase in abundance when NMD is inhibited than those with 3' UTRs shorter than 400nts (KS test, $p = 3.84 \times 10^{-10}$, D = 0.06); however, a PTC_{50nt} has a 4-fold stronger effect even for transcripts with 3'UTR < 400 nts (KS test, $p < 2 \times 10^{-308}$, D = 0.23). We similarly investigated the effects of CDS length, for which there is some evidence that a longer CDS may enhance NMD (for NTC-containing isoforms, CDS < 400nts vs > 2000nts: KS test, $p < 1.58 \times 10^{-7}$, D = 0.07), and GC content of the 3' UTR, for which higher GC content appears to strongly increase likelihood of NMD (NTC-containing isoforms, 3' UTR %GC < 35% vs > 55%: $p < 2 \times 10^{-308}$, D = 0.18).

Premature termination codons generated by alternative splicing events or uORFS

We also explored the types of molecular events that can generate a putative PTC, including alternative splicing and upstream open reading frames (uORFs). Since all of the transcripts in our high-confidence set of NMD targets come from genes that also produce NTC-containing isoforms, we infer that the PTCs were generated by an alternative splicing event. Using the JuncBASE program[75] to investigate these alternative splicing events, we found that 365 of the isoforms (13%, of 2,793) were generated by the introduction of a PTC-containing cassette exon (poisonous exon) and 54 isoforms had a retained intron containing a PTC. Interestingly, most of the transcripts (2,117, 75.8%) have an alternative splicing event that causes a frameshift generating a downstream PTC. Additionally, 257 NMD-targeted isoforms (9%) were generated by the introduction of an intron downstream of the termination codon of the productive coding sequence. One interesting example of this is *KHDRBS1*, encoding a protein involved in both signal transduction and mRNA processing, for which alternative splicing combined with alternative poly-adenylation results in the formation of an NMD target without affecting the structure of the coding sequence (Figure 2.4C).

NMD may also act on transcripts with translated uORFs because the uORF termination codon will almost always be recognized as premature, provided that the main coding sequence is not also translated. We examined NTC-containing transcripts for uORFs. We found that 14,348 NTC-containing transcripts had at least one uORF, but only 5,628 of these had a uORF with a strong Kozak signal sequence for translation initiation[76].

Table 2.5 Features of uORF-containing transcripts that are degraded by NMD

			FPKM >1	Increased abundance	Decreased abundance	Ratio
uORF: strong Kozak signal	uORF overlaps main CDS		408	165 (40%)	90 (22%)	+1.83
	No overlap	uORF ≥35aa long	602	269 (45%)	119 (20%)	+2.26
		uORF <35aa long	4,618	1,661 (36%)	1,054 (23%)	+1.58
uORF: weak Kozak signal	uORF overlaps main CDS		3,178	996 (31%)	891 (28%)	+1.12
	No overlap	uORF ≥35aa long	1,425	437 (31%)	341 (24%)	+1.28
		uORF <35aa long	4,117	1,113 (27%)	1,192 (29%)	-1.07
Transcripts with no uORF			9,540	2,244 (23%)	3,041 (32%)	-1.36
All NTC-containing transcripts			23,888	6,885 (29%)	6,728 (28%)	+1.02

Increased and decreased abundance refers to the expression changes in NMD-inhibited cells (compared to control cells) that are significant according to Cuffdiff and at least >1.5 fold. The percentage is of the number of transcripts in that category. 'weak' and 'strong' refer to the strength of the Kozak signal sequence around the uORF start codon. A positive ratio (+) is increased/decreased and a negative ratio (-) is decreased/increased.

Transcripts with a 'strong' uORF that either overlaps the main CDS or is at least 35aa long are about twice as likely to significantly increase in abundance when NMD is inhibited, compared to those with no uORF (Table 2.5). Transcripts with a shorter 'strong' uORF are also more likely to increase, as are those with 'weak' uORFs. Thus, our data suggests that hundreds to thousands of transcripts with uORFs are naturally targets of NMD and could be regulated through the potential translation of uORFs.

Discussion

Identification of novel isoforms by RNA-Seq enabled discovery of thousands of NMD targets

In this study, through depletion of UPF1 (a key regulator in the NMD pathway), we stabilized transcripts that are otherwise degraded by NMD and determined their structure and expression via RNA-seq analysis. We report a set of 2,793 transcripts that contain a premature termination codon according to the 50nt rule and are stabilized by inhibition of the NMD pathway. The high abundance of these transcripts in the NMD-inhibited sample indicates that they are unlikely to be produced by random errors. Our results confirmed most well-known NMD-targets, such as those derived from the SR genes[6,7], and discovered many more derived from other genes, including many splicing factors. Numerous genes involved in a variety of other functional categories including aminoacyl-tRNA biosynthesis, proteolysis, and metabolic pathways, were also found to produce NMD-targeted transcripts, suggesting that a broad range of biological processes may be affected by NMD. Transcripts degraded by NMD are significantly enriched for ultraconserved elements of the

human genome, supporting a potential role for these elements in regulation through alternative splicing-coupled NMD. Ultimately, we extrapolate that 20% or more of human genes produce NMD-targeted alternative isoforms and are potentially post-transcriptionally regulated by NMD coupled with alternative splicing.

Many NMD-targeted genes are thought to be auto-regulated through a feedback loop wherein protein abundance affects the relative levels of productive and unproductive transcripts. The model is that the protein abundance directly or indirectly affects splicing of its own pre-mRNA so that when a protein is present at an unnecessarily high level, the splicing apparatus shifts to generating NMD-targeted isoforms. We found that when NMD is inhibited, over 500 NMD-targeted genes appear to shift splicing toward generating the NMD-targeted isoform, while at the same time down-regulating the expression of NTC-containing isoforms, which likely results in the lower abundance of those proteins. This group of genes includes many involved in RNA splicing. In fact, over half (18/31) of the NMD-targeted genes with the GO annotation 'RNA processing' fall into this group, as do all 6 genes annotated as 'heterogeneous nuclear ribonucleoprotein complex'. We also found 180 NMD-targeted genes with decreased expression at both the transcriptional level and the NTC-containing isoform level. These are enriched for genes involved in the ribosome and translation, and their down-regulated expression might be a side effect of the UPF1 knockdown, which results in slowed cell growth compared to control cells. Such down-regulation may be mediated through the concerted effort of transcriptional regulation and alternative splicing coupled with NMD. Thus hundreds of genes were found with evidence of possible regulation through NMD, indicating such a regulatory system may be much more prevalent than previous thought.

The 50nt rule is the strongest predictor of NMD in human cells, although other transcript features may be important

While the 50nt rule has been broadly believed to be the general signal for NMD degradation in mammals[46], a longer 3'-UTR has been proposed to also play a role[50,52,55]. Results indicating that a long 3' UTR is enough to cause NMD have come from experiments done with a particular construct, or by looking at the annotated 3' UTR length of genes or transcripts that increase in overall abundance when NMD is inhibited. In our present data, we identify the precise isoforms that are degraded by NMD in normal cells, thereby allowing for the discovery of novel instances of PTC_{50nt} as well as directly measure the 3' UTR length of a specific isoform. Strikingly, the average isoform abundance change when NMD is inhibited is a 3.3-fold increase for PTC_{50nt}-containing transcripts, compared to no change on average for NTC-containing transcripts. Since we find that the majority of PTC_{50nt} fall in the middle of the main coding region, those transcripts are going to have longer 3' UTRs, but a long 3' UTR in the absence of a PTC_{50nt} has a much smaller effect on isoform abundance. This lack of a strong global effect by 3' UTR length agrees with reports of elements in the 3' UTR, near the termination codon, that protect against NMD[77], including the binding of PTBP1 [78]. Of the 30% of transcripts with a PTC_{50nt} that do not increase at least 1.2-fold when NMD is inhibited, it is possible that as little as 5%of those are actually escaping NMD. For the others we see evidence of potentially confounding transcriptional or splicing regulation in the gene due to secondary effects of the NMD inhibition (including transcriptional down-regulation or a shift in splicing toward the

productive isoform) and of uncertainty in the analysis (such as low sequencing coverage, incorrect transcript assembly or CDS definition, and complex alternative splicing patterns). PTC_{50nt}-containing transcripts can escape NMD through secondary structure in the 3' UTR[54,79,80], alternative poly-adenylation sites, or differential deposition of the exon junction complex required for NMD[81,82]. We conclude that while 3' UTR length may have an effect for a subset of genes, the 50nt rule is likely the major mechanism of targeting a transcript for NMD in humans.

Based on the above, we would not expect transcripts without a PTC_{50nt} to be degraded by NMD. While we find that almost 30% of NTC-containing transcripts significantly increase when NMD is inhibited, almost the same number decrease in abundance, leading us to believe that these changes are likely due to secondary effects of the knockdown experiment. Despite this, we sought features of these increasing transcripts that may indicate they are targeted for NMD. As described above, we find little evidence that 3' UTR length significantly affects likelihood of degradation. The length of the coding sequence also does not have a strong effect. We noticed a correlation between 3' UTR GC content and likelihood of increased abundance that warrants further exploration, particularly since it has been shown that UPF1 binds G-rich sequences[52]. Finally, we found that for thousands of transcripts, having a uORF with a strong Kozak signal sequence almost doubles the likelihood that they increase in abundance in UPF1-depleted cells, indicating that NMD may also be involved in extensive regulation of gene expression via a mechanisms that is sensitive to the translation of uORFs.

Materials and Methods

Knockdown of UPF1 by shRNA

HeLa cells were inoculated on plates in Dulbecco's MEM medium with 0.1 mM non-essential amino acids and 10% fetal bovine serum, incubated at 37°C and 5% CO_2. Plates at 80% cell confluency were transfected with plasmids pSUPERpuro-hUpf1/II and pSUPERpuro-Scramble (a gift from Oliver Mühlemann's lab[62,63]), whose functions were to knockdown UPF1 and act as a mock control, respectively. Transfections were performed using Lipofectamine™ LTX and PLUS™ Reagents (Invitrogen) according to the manufacturer's protocol, and the following culture according to the published method[63]. The whole cell lysates were prepared in 1% sodium dodecyl sulfate and incubated at 100 °C for 5 min, then centrifuged at 12,000 rpm for 10 min. The extracted total protein was quantified using the micro BCA™ protein assay kit (Pierce Company). Knockdown efficiency was validated by Western blot analysis, using β-actin as the control. Odyssey® infrared imaging system (Li-Cor) was used to quantify and compare the UPF1 protein level between UPF1-knockdown and control samples. Total RNA was extracted using the QIAGEN RNeasy® Mini kit according to the manufacturer's manual; RNA concentration was determined by NanoDrop 2000®, and RNA integrity was determined by 1.2% agar gel and BioAnalyzer. NMD inhibition was validated by amplification of several known NMD-degraded transcripts by real-time PCR with ABI 7500 fast real time PCR system (Applied Biosystems), and data were analyzed using 7500 software v2.04 (Applied Biosystems).

Preparation of RNA-Seq libraries

Directional and paired-end RNA-seq libraries were constructed according to the protocol published on the Illumina website (www.illumina.com), with a few changes: The adapters were prepared according to the reported methods[83]. The PCR process to prepare the library was divided in two steps. In the first step, 3 cycles of PCR were performed (according to the protocol) to prepare the library template, and then the library was run on a 2% agarose gel, fragments of desired size were cut out and isolated by QIAquick® Gel Extraction Kit. A second round of PCR (12 cycles) was performed to enrich the library and then it was purified twice with Agencourt AMPure XP kit as suggested by the Illumina protocol. The libraries were then assayed by Agilent 2100 BioAnalyzer. These RNA-Seq libraries were prepared from cells with inhibited NMD and control cells for two biological replicates. One biological replicate was sequenced on an Illumina GAIIx machine and the other on HiSeq 2000.

Transcript assembly and abundance quantification

Paired end reads for each library were aligned to the NCBI human RefSeq transcriptome[84] with Bowtie[85] to determine the average insert size and standard deviation, required as a parameter by TopHat [64]. The reads of each library were then aligned to the human genome (hg19 assembly, Feb. 2009) using TopHat v1.2.0 with default parameters plus the following: --coverage search, --allow indels, --microexon search, and --butterfly search. Cufflinks 1.0.1[65,66] was used to assemble each set of aligned reads into transcripts with the UCSC known transcript set (http://genome.ucsc.edu/,[86]) as the reference guide, along with the following parameters: --frag-bias-correct, and --multi-read-correct. Cuffcompare (a sub-tool of Cufflinks) was used to merge the resulting sets of assembled transcripts. Each junction was assigned a Shannon entropy score based on offset and depth of spliced reads across all four libraries. Transcripts with a junction that had an entropy score <1 and was not present in the reference annotation were filtered out. Cuffdiff (a sub-tool of Cufflinks) was used to quantify and compare transcript abundance (measured by FPKM, Fragments Per Kilobase per Million reads) between the UPF1 knockdown and control samples. For each sample, the reads from two biological replicates were provided. The following parameters were used: --frag-bias-correct and --multi-read-correct. Only transcripts with FPKM>1 in either the control or UPF1 knockdown sample were used for further analysis. A transcript was called significantly more abundant in the UPF1 knockdown sample if Cuffdiff called it significantly changing and the fold change was greater than 1.5x. Significantly decreased transcript abundances were determined in the same way.

Determination of NMD targets

For each transcript, the coding sequence (CDS) was determined as described in the Supplementary Methods. A coding sequence was defined to terminate in a premature stop codon (PTC_{50nt}) if it stops at least 50 nucleotides upstream of the last exon-exon junction (50nt rule in mammals). NMD targets were defined as those transcripts with both a PTC_{50nt} and significantly increased expression abundance in NMD inhibited (UPF1 knockdown) cells. The transcripts must also increase in each biological replicate when analyzed independently and come from a gene with a NTC-containing isoform with FPKM>0. To obtain a more reliable list of NMD-targeted transcripts, only those transcripts that adhered

to either of the following criteria were kept: 1) No NTC-containing isoform from the gene was more than 1.2-fold higher in the NMD inhibited sample, or 2) the PTC_{50nt}-containing isoform increased at least 2x more than the sum of all NTC-containing isoform FPKMs from the gene in NMD inhibited cells. Functional classification was based on Ensemble gene annotation, using Gene Ontology (GO) terms[73]. Alternative splicing events were characterized with JuncBASE[75].

Validation by real-time PCR

To validate the potential NMD-targeted transcripts found in the present study, we measured their relative mRNA abundance in UPF1-KD and the mock control cells using quantitative real-time PCR. Total RNAs were extracted as described above. cDNAs were synthesized from 5 μg total RNA using the SuperScript II first-strand cDNA synthesis system (Invitrogen) according to the manufacturer's instructions. The isoform-specific primers used are listed in Table S4. Real-time PCRs were performed on Applied Biosystems 7500 Fast Real-Time PCR System using the respective pair of primers designed with Primer Premier 6.0 and Maxima SYBR Green/ROX qPCR Master Mix (Thermo Scientific). PCR reactions were performed in four replicates, and the expression levels were normalized to that of β-actin control gene (NM_001101).

References

1. Kervestin, S. & Jacobson, A. NMD: a multifaceted response to premature translational termination. *Nat. Rev. Mol. Cell Biol.* **13,** 700–712 (2012).
2. Isken, O. & Maquat, L. E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev.* **21,** 1833–1856 (2007).
3. Lykke-Andersen, J. & Bennett, E. J. Protecting the proteome: eukaryotic cotranslational quality control pathways. *J. Cell Biol.* (2014).
4. He, F., Peltz, S. W., Donahue, J. L., Rosbash, M. & Jacobson, A. Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1- mutant. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 7034–7038 (1993).
5. Pulak, R. & Anderson, P. mRNA surveillance by the Caenorhabditis elegans smg genes. *Genes Dev.* (1993).
6. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446,** 926–929 (2007).
7. Ni, J. Z. *et al.* Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* **21,** 708–718 (2007).
8. Mudge, J. M. *et al.* The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.* **28,** 2949–2959 (2011).
9. Yan, Q. *et al.* Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 3445–3450 (2015).
10. Lareau, L. F. & Brenner, S. E. Regulation of Splicing Factors by Alternative Splicing

and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. *Mol. Biol. Evol.* **32,** 1072–1079 (2015).
11. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 189–192 (2003).
12. Rehwinkel, J., Letunic, I., Raes, J., Bork, P. & Izaurralde, E. Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA* **11,** 1530–1544 (2005).
13. Schweingruber, C., Rufener, S. C., Zünd, D., Yamashita, A. & Mühlemann, O. Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells. *Biochim. Biophys. Acta* **1829,** 612–623 (2013).
14. He, F. & Jacobson, A. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annu. Rev. Genet.* (2015). doi:10.1146/annurev-genet-112414-054639
15. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* **16,** 665–677 (2015).
16. Jumaa, H. & Nielsen, P. J. The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J.* **16,** 5077–5085 (1997).
17. Sureau, A., Gattoni, R., Dooghe, Y., Stévenin, J. & Soret, J. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J.* **20,** 1785–1796 (2001).
18. Spellman, R., Llorian, M. & Smith, C. W. J. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol. Cell* **27,** 420–434 (2007).
19. Rossbach, O. *et al.* Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. *Mol. Cell. Biol.* **29,** 1442–1451 (2009).
20. Sun, S., Zhang, Z., Sinha, R., Karni, R. & Krainer, A. R. SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nat Struct Mol Biol* **17,** 306–312 (2010).
21. McGlincy, N. J. *et al.* Expression proteomics of UPF1 knockdown in HeLa cells reveals autoregulation of hnRNP A2/B1 mediated by alternative splicing resulting in nonsense-mediated mRNA decay. *BMC Genomics* **11,** 565 (2010).
22. Dredge, B. K. & Jensen, K. B. NeuN/Rbfox3 nuclear and cytoplasmic isoforms differentially regulate alternative splicing and nonsense-mediated decay of Rbfox2. *PLoS ONE* **6,** e21585 (2011).
23. Anko, M.-L. *et al.* The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol.* **13,** R17 (2012).
24. Jangi, M., Boutz, P. L., Paul, P. & Sharp, P. A. Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev.* **28,** 637–651 (2014).
25. Saltzman, A. L. *et al.* Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol. Cell. Biol.* **28,** 4320–4330 (2008).
26. Lareau, L. F., Brooks, A. N., Soergel, D. A. W., Meng, Q. & Brenner, S. E. The coupling of alternative splicing and nonsense-mediated mRNA decay. *Adv. Exp. Med. Biol.* **623,** 190–211 (2007).

27. Hyvönen, M. T. *et al.* Polyamine-regulated unproductive splicing and translation of spermidine/spermine N1-acetyltransferase. *RNA* **12,** 1569–1582 (2006).
28. Hyvönen, M. T. *et al.* Tissue-specific alternative splicing of spermidine/spermine N (1)-acetyltransferase. *Amino Acids* **42,** 485–493 (2012).
29. Bhattacharya, A. *et al.* Characterization of the biochemical properties of the human Upf1 gene product that is involved in nonsense-mediated mRNA decay. *RNA* **6,** 1226–1235 (2000).
30. Lykke-Andersen, J., Shu, M.-D. & Steitz, J. A. Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. *Cell* **103,** 1121–1131 (2000).
31. Gehring, N. H. *et al.* Exon-junction complex components specify distinct routes of nonsense-mediated mRNA decay with differential cofactor requirements. *Mol. Cell* **20,** 65–75 (2005).
32. Kashima, I. *et al.* Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes Dev.* **20,** 355–367 (2006).
33. Ivanov, P. V., Gehring, N. H., Kunz, J. B., Hentze, M. W. & Kulozik, A. E. Interactions between UPF1, eRFs, PABP and the exon junction complex suggest an integrated model for mammalian NMD pathways. *EMBO J.* **27,** 736–747 (2008).
34. Huntzinger, E., Kashima, I., Fauser, M., Saulière, J. & Izaurralde, E. SMG6 is the catalytic endonuclease that cleaves mRNAs containing nonsense codons in metazoan. *RNA* **14,** 2609–2617 (2008).
35. Yamashita, A. *et al.* SMG-8 and SMG-9, two novel subunits of the SMG-1 complex, regulate remodeling of the mRNA surveillance complex during nonsense-mediated mRNA decay. *Genes Dev.* **23,** 1091–1105 (2009).
36. Melero, R. *et al.* Structures of SMG1-UPFs Complexes: SMG1 Contributes to Regulate UPF2-Dependent Activation of UPF1 in NMD. *Structure* **22,** 1105–1119 (2014).
37. Deniaud, A. *et al.* A network of SMG-8, SMG-9 and SMG-1 C-terminal insertion domain regulates UPF1 substrate recruitment and phosphorylation. *Nucleic Acids Res.* **43,** 7600–7611 (2015).
38. Lykke-Andersen, S. *et al.* Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes Dev.* **28,** 2498–2517 (2014).
39. Hwang, J., Sato, H., Tang, Y., Matsuda, D. & Maquat, L. E. UPF1 association with the cap-binding protein, CBP80, promotes nonsense-mediated mRNA decay at two distinct steps. *Mol. Cell* **39,** 396–409 (2010).
40. Chakrabarti, S. *et al.* Molecular mechanisms for the RNA-dependent ATPase activity of Upf1 and its regulation by Upf2. *Mol. Cell* **41,** 693–703 (2011).
41. Okada-Katsuhata, Y. *et al.* N- and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD. *Nucleic Acids Res.* **40,** 1251–1266 (2012).
42. Lasalde, C. *et al.* Identification and functional analysis of novel phosphorylation sites in the RNA surveillance protein Upf1. *Nucleic Acids Res.* **42,** 1916–1929 (2014).
43. Nicholson, P., Josi, C., Kurosawa, H., Yamashita, A. & Mühlemann, O. A novel phosphorylation-independent interaction between SMG6 and UPF1 is essential for human NMD. *Nucleic Acids Res.* **42,** 9217–9235 (2014).

44. Chakrabarti, S., Bonneau, F., Schüssler, S., Eppinger, E. & Conti, E. Phospho-dependent and phospho-independent interactions of the helicase UPF1 with the NMD factors SMG5-SMG7 and SMG6. *Nucleic Acids Res.* **42,** 9447–9460 (2014).
45. Kerényi, Z. *et al.* Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO J.* **27,** 1585–1595 (2008).
46. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* **23,** 198–199 (1998).
47. Zhang, J., Sun, X., Qian, Y. & Maquat, L. E. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA* **4,** 801–815 (1998).
48. Amrani, N. *et al.* A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature* **432,** 112–118 (2004).
49. Singh, G., Rebbapragada, I. & Lykke-Andersen, J. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol.* **6,** e111 (2008).
50. Yepiskoposyan, H., Aeschimann, F., Nilsson, D., Okoniewski, M. & Mühlemann, O. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. *RNA* **17,** 2108–2118 (2011).
51. Kervestin, S., Li, C., Buckingham, R. & Jacobson, A. Testing the faux-UTR model for NMD: Analysis of Upf1p and Pab1p competition for binding to eRF3/Sup35p. *Biochimie* (2012). doi:10.1016/j.biochi.2011.12.021
52. Hurt, J. A., Robertson, A. D. & Burge, C. B. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.* **23,** 1636–1650 (2013).
53. Kurosaki, T. & Maquat, L. E. Rules that govern UPF1 binding to mRNA 3' UTRs. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 3357–3362 (2013).
54. Behm-Ansmant, I., Gatfield, D., Rehwinkel, J., Hilgers, V. & Izaurralde, E. A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *EMBO J.* **26,** 1591–1601 (2007).
55. Hogg, J. R. & Goff, S. P. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell* **143,** 379–389 (2010).
56. Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F. & Dietz, H. C. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat. Genet.* **36,** 1073–1078 (2004).
57. Pan, Q. *et al.* Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* **20,** 153–158 (2006).
58. Sayani, S., Janis, M., Lee, C. Y., Toesca, I. & Chanfreau, G. F. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Mol. Cell* **31,** 360–370 (2008).
59. Hansen, K. D. *et al.* Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. *PLoS Genet.* **5,** e1000525 (2009).
60. Weischenfeldt, J. *et al.* Mammalian tissues defective in nonsense-mediated mRNA

decay display highly aberrant splicing patterns. *Genome Biol.* **13,** R35 (2012).
61. Tani, H. *et al.* Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability. *RNA Biol* **9,** 1370–1379 (2012).
62. Bühler, M., Steiner, S., Mohn, F., Paillusson, A. & Mühlemann, O. EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length. *Nat Struct Mol Biol* **13,** 462–464 (2006).
63. Paillusson, A., Hirschi, N., Vallan, C., Azzalin, C. M. & Mühlemann, O. A GFP-based reporter system to monitor nonsense-mediated mRNA decay. *Nucleic Acids Res.* **33,** e54 (2005).
64. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
65. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
66. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
67. Kim, Y. K., Furic, L., DesGroseillers, L. & Maquat, L. E. Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. *Cell* **120,** 195–208 (2005).
68. Kaygun, H. & Marzluff, W. F. Regulated degradation of replication-dependent histone mRNAs requires both ATR and Upf1. *Nat Struct Mol Biol* **12,** 794–800 (2005).
69. Taylor, M. S. *et al.* Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* **155,** 1034–1048 (2013).
70. Moriarty, P. M., Reddy, C. C. & Maquat, L. E. Selenium deficiency reduces the abundance of mRNA for Se-dependent glutathione peroxidase 1 by a UGA-dependent mechanism likely to be nonsense codon-mediated decay of cytoplasmic mRNA. *Mol. Cell. Biol.* **18,** 2932–2939 (1998).
71. Seyedali, A. & Berry, M. J. Nonsense-mediated decay factors are involved in the regulation of selenoprotein mRNA levels during selenium deficiency. *RNA* **20,** 1248–1256 (2014).
72. Huang, L. *et al.* RNA homeostasis governed by cell type-specific and branched feedback loops acting on NMD. *Mol. Cell* **43,** 950–961 (2011).
73. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
74. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
75. Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.* **21,** 193–202 (2011).
76. Kozak, M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* **266,** 19867–19870 (1991).
77. Toma, K. G., Rebbapragada, I., Durand, S. & Lykke-Andersen, J. Identification of elements in human long 3' UTRs that inhibit nonsense-mediated decay. *RNA* **21,** 887–897 (2015).
78. Ge, Z., Quek, B. L., Beemon, K. L. & Hogg, J. R. Polypyrimidine tract binding protein 1 protects mRNAs from recognition by the nonsense-mediated mRNA decay pathway. *Elife* **5,** (2016).

79. Eberle, A. B., Stalder, L., Mathys, H., Orozco, R. Z. & Mühlemann, O. Posttranscriptional gene regulation by spatial rearrangement of the 3' untranslated region. *PLoS Biol.* **6,** e92 (2008).
80. Isken, O. & Maquat, L. E. The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat. Rev. Genet.* **9,** 699–712 (2008).
81. Saulière, J. *et al.* The exon junction complex differentially marks spliced junctions. *Nat Struct Mol Biol* **17,** 1269–1271 (2010).
82. Alexandrov, A., Colognori, D., Shu, M.-D. & Steitz, J. A. Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 21313–21318 (2012).
83. Vigneault, F., Sismour, A. M. & Church, G. M. Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. *Nat. Methods* **5,** 777–779 (2008).
84. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37,** D32–6 (2009).
85. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
86. Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39,** D876–82 (2011).

Chapter 3 Evidence for pervasive regulation via nonsense-mediated mRNA decay across several diverse species

Abstract

To investigate the prevalence of potential regulation via nonsense-mediated mRNA decay, particularly when coupled to alternative splicing, in a number of diverse eukaryotic species, we have analyzed RNA-Seq data from cells where NMD has been inhibited via depletion of UPF1, a critical and conserved protein in the degradation pathway. We found that hundreds to thousands of genes produce alternative isoforms that are potentially degraded by NMD in each of the species studied (human, mouse, frog, zebrafish, fly, *Arabidopsis*, and *S. pombe*). We also found that an exon-exon junction over 50 nucleotides downstream of the termination codon is a strong predictor of NMD degradation in human cells, and also seems to play a role in the other species tested, with the exclusion of *S. pombe*. In contrast, we found little to no correlation between the likelihood of degradation by NMD and 3' UTR length in any of the species, except perhaps for *S. pombe*. Additionally, we found that hundreds of genes per species produce transcripts with a uORF that could potentially trigger NMD. Overall, we found that regulation via alternative splicing coupled with NMD has a substantial effect on the transcriptome in each species tested including animals, plants, and fungi.

Introduction

Nonsense-mediated mRNA decay (NMD) is a surveillance pathway that degrades mRNA transcripts containing a premature or aberrant termination codon. This process prevents the production of potentially harmful truncated proteins from transcripts with transcription or splicing errors or from transcripts from genes that have acquired a nonsense mutation[1-4]. Many NMD factors, in particular the core NMD factor UPF1, are conserved throughout the major branches of the eukaryotic lineage indicating the importance of this pathway for plants, fungi, and animals[5-9]. In addition to a role in cellular surveillance, NMD has also been shown to be important in post-transcriptional gene regulation, often but not always when coupled with alternative splicing[10-14].

While some NMD factors, including UPF1, also known as SMG2 (*C. elegans*) or RENT1 (*S. cerevisiae*), are conserved throughout eukaryotes, the mechanism of NMD varies between species. In mammalian cells, UPF1 is recruited to a premature termination codon and phosphorylated by SMG1. Phosphorylated UPF1 then recruits the factors needed for SMG6-dependent endonucleoytic degradation and SMG7-dependent deadenylation and decapping (reviewed in[13,14]). There are few different models for how UPF1 is recruited and NMD is triggered. For one, the presence of an exon-exon junction, marked with an exon junction complex, over 50 nucleotides downstream of the termination codon recruits UPF1 (the '50nt Rule'[15,16]). Alternatively, a long 3' UTR has been proposed to trigger NMD either by leading to aberrant termination (i.e. via the increased distance between the poly-A binding protein and terminating ribosome)[3,17] or through the length-dependent binding of UPF1[18].

The mechanism of NMD appears to be similar throughout vertebrates: the NMD-inducing effects of a downstream exon-exon junction and the requirement for SMG6 have been shown in zebrafish[7]. In contrast, for fly and worm, most reports indicate that 3' UTR length is the primary effector[17,19-21]. In yeast, which does not have SMG1, SMG5, SMG6, or SMG7, 3' UTR structural feature including length is thought to trigger NMD (*faux*-3' UTR model) and degradation happens via decapping (reviewed in[22]). Plant NMD relies on a SMG7-dependent exonucleolytic pathway and both a long 3' UTR and a downstream exon-exon junction have been shown to trigger it[23] (reviewed in[24]). Additionally, there is evidence that the presence of an upstream open reading frame (uORF) can elicit NMD in mammals[25], worm[26], yeast[22], and plants[27].

The role of NMD in post-transcriptional gene regulation, particularly when coupled to alternative splicing, has been investigated in a number of species. Studies in human have revealed that 20% of alternatively spliced genes are putative NMD targets and these are enriched for splicing factors[28,29] (see Chapter 1). Hundreds of NMD-targeted genes (up to 15-20%) have also been reported in mouse[25,30] and fly[11,19,20]. In plants, 15-18% of multi-exonic genes have been reported as putative NMD targets[31]. NMD has also been shown to play a key role in differentiation and development in mammals, frog, zebrafish, worm, and fly[7,21,32-35] and in stress responses in human, yeast[36,37], and plants[38].

While the NMD pathway is conserved throughout eukaryotes, its effect on organism viability varies from species to species. NMD deficiency is lethal in mammals, zebrafish, fly, and plants, but not in worm and yeast. The cause of this lethality could be either the overall accumulation of aberrant PTC-containing transcripts or the specific mis-regulation of a gene or set of genes. In the plant, *Arabidopsis thaliana*, the lethality is due to hyperactivation of the defense response pathway[39]. Recent work has indicated that the mis-regulation of the apoptosis-inducing gene, Gadd45, is a major source of lethality upon NMD impairment in mammals and fly[40]. There are other examples of the conservation of NMD targets between species including neuron-specific Psd95 which is subject to regulation via alternative splicing coupled to NMD in mammals but not invertebrates[41], SRSF5 for which the NMD-inducing alternative splice event is conserved throughout animals and even seen in fungal species[42], and the NMD factors themselves in a feedback loop conserved in human, zebrafish, and worm[43].

In this study, we analyzed published and unpublished RNA-seq data sets from UPF1-depletion experiments in human, mouse, frog, zebrafish, fly, yeast, and plant. Systematically analyzing the data in analogous ways across the different species allowed us to investigate the similarities and differences in the effects that NMD has on the transcriptomes of these diverse species. We find that a substantial fraction of genes have the potential to be regulated via alternative splicing coupled with NMD in each species. Additionally, uORFs potentially have a strong effect on NMD susceptibility in most species. We also gained insight into the mechanisms triggering NMD: all species investigated, except for *S. pombe*, appear to be susceptible to some extent to the 50nt Rule while 3' UTR length does not have a strong effect.

Results

Analysis of RNA-seq data from UPF1-depletion studies for seven eukaryotic species

In order to investigate the effect of NMD on the transcriptomes of diverse eukaryotic species, we used RNA-seq data from experiments where NMD has been inhibited via the depletion of UPF1, a core and conserved NMD factor. We generated the data sets for human, frog, zebrafish, *Arabidopsis*, and *S. pombe*[*] and used published data for mouse[25] and fly[44]. For human HeLa cells, mouse ES cells, fly S2 cells, and *Arabidopsis* plants, UPF1 mRNA levels were decreased by 47-82% (Table 3.1) via shRNA, RNAi, or genetic insertion (Table 3.2). For frog and zebrafish, UPF1 protein expression was perturbed in embryos by translation- and/or splice-blocking morpholinos (Table 3.2). A strain with a UPF1 deletion was used for *S. pombe*. Two to four biological replicates of each condition were sequenced with Illumina technology; RNA-seq library preparation protocols and read length, type, and depth vary between species (summarized in Table 3.2).

For each species, reads were aligned to the appropriate genome with Tophat (v2.1.0) and known and novel transcripts were assembled and quantified with Cufflinks (v2.2.1). In order to determine which transcripts have a potential premature termination codon (PTC), we predicted the coding sequence (CDS) of each transcript. We also focused on just those transcripts that had an FPKM (fragments per kilobase, normalized for fragment depth) value greater than 1 in at least one of the two conditions (control and NMD-inhibited) since differential expression of transcripts that are very low expressed in both are more subject to experimental noise.

Depletion of UPF1 and the resulting inhibition of NMD affect the abundance of 9-45% of transcripts and 2-37% of genes for each species (Table 3.3). The amount of overall changes to the transcriptomes after UPF1-depletion varies between species and is indicative of both biological and experimental differences. For human and frog, almost half

Table 3.1 UPF1 mRNA depletion efficiency by experiment

Species	UPF1 control FPKM	UPF1 experimental FPKM	UPF1 mRNA % remaining
Homo sapiens	33.61	6.03	18%
Mus musculus	38.21	9.09	24%
Drosophila melanogaster	30.61	16.26	53%
Arabidopsis thaliana	34.27	9.37	27%
Schizosaccharomyces pombe[$]	89.41	0	0%

[$]For *S. pombe*, the experimental sample is a UPF1 deletion strain and no mRNA is expected.
FPKM: fragments per kilobase, normalized for fragment depth

[*] The human cell line experiments were performed by Anna Desai and the *Arabidopsis* experiments by Gang Wei, both in Steven Brenner's lab. The frog experiments were performed by Darwin S. Dichmann in Richard M. Harland's lab. The zebrafish experiments were performed by Thomas L. Gallagher in Sharon L. Amacher's lab. The *S. pombe* experiments were performed by Maki Inada.

Table 3.2 Summary of UPF1-depletion RNA-seq experiments

Species	Cell line/tissue	NMD inhibition	Number of replicates	Library and read type	Number of read per sample	Number of aligned reads per sample
Homo sapiens	HeLa cells	shRNA knockdown of UPF1	2x control 2x UPF1 KD	Poly-A selected, stranded library; 80bp or 101bp, PE reads	42-182M (21-91M fragments)	37-136M (75-90% mapping rate)
Mus musculus	ES cells	shRNA knockdown of UPF1	2x control 4x UPF1 KD	Poly-A selected, stranded library; 80bp, PE reads	24-64M (12-32M fragments)	20-56M (81-89% mapping rate)
Xenopus tropicalis	Embryo (stage 14)	Morpholinos against UPF1	3x control 2x anti-translation MO 1x splice blocking MO	Poly-A selected, stranded library; 250bp, PE reads	5.5-11M (2.5-5.5M fragments)	3.8M-6.4M (58-69% mapping rate)
Danio rerio	Embryo (11hr)	Morpholinos against UPF1	2x control 2x splice blocking MO	Poly-A selected, unstranded library; 100bp, PE reads	106-131M (53-66M fragments)	79-100M (74-77% mapping rate)
Drosophila melanogaster	S2-DRSC cells	RNAi knockdown of UPF1	2x control 2x UPF1 KD	Poly-A selected, unstranded library; 80bp, SE reads	17-22M (17-22M fragments)	14-20M (67-89% mapping rate)
Arabidopsis thaliana	seedling	UPF1 mutant (knockdown) *upf1-5*	3x wild-type 3x UPF1 mutant	Poly-A selected, stranded library; 250bp, PE reads	7-15M (3.5-8M fragments)	7-12M (51-84% mapping rate)
Schizosaccharomyces pombe	NA	UPF1 mutant (deletion)	2x wild-type 2x UPF1 mutant	Poly-A selected, stranded library; 250bp, PE reads	6.5-9M (3-4.5M fragments)	5-7.5M (82-83% mapping rate)

of the substantially expressed CDS$^+$ transcripts change over 2-fold in abundance upon NMD inhibition. In contrast, 20% or fewer transcripts change for the other species and the number is particularly low for zebrafish and *S. pombe*. However, for all species there is at least a slight bias toward increased abundance over decreased abundance as is expected for inhibiting a degradation pathway (direct effects are the increases in abundance of no longer degraded transcripts). The reasons that we observe such variation in transcriptome changes between the species include those both biological (UPF1 and/or NMD have different influences in a given species or cell line/tissue) and experimental (efficiency of NMD inhibition may vary between experiments, lower read depth limits our ability to measure expression changes, etc).

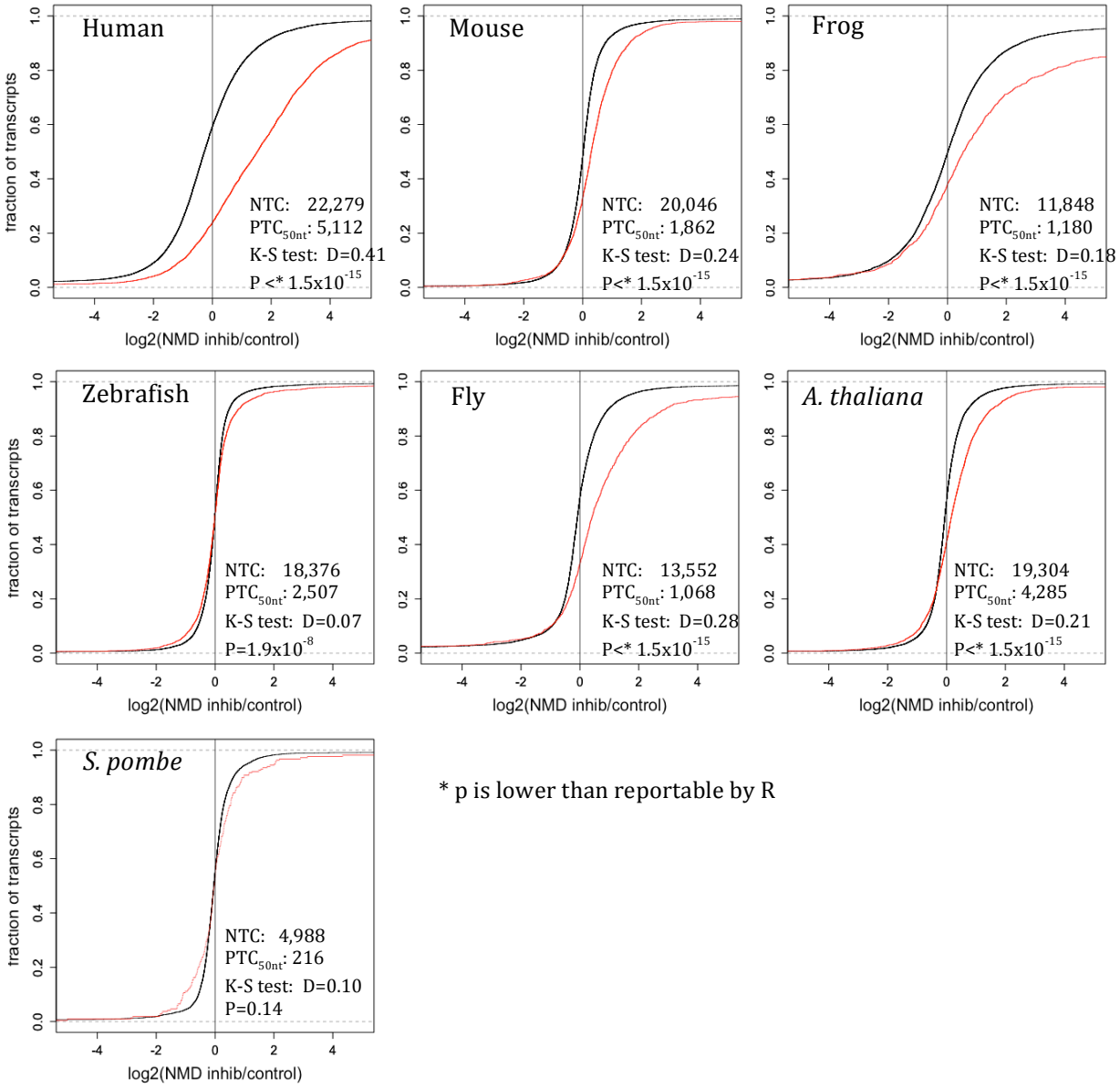
Features of NMD-affected transcripts

With this data, we are able to investigate how features expected to trigger NMD affect a transcript's susceptibility to being affected by depletion of UPF1. First we looked at the '50nt Rule' by determining which transcripts contain a termination codon more than 50 nucleotides upstream of the last exon-exon junction of the transcript (PTC_{50nt}). We compared the distribution of the $\log_2$(NMD inhibition FPKM/control FPKM) values for this set of transcripts to that of the set of non-PTC_{50nt}-containing transcripts (NTC transcripts). The expectation is that if a feature triggers NMD, the distribution should be shifted to the right because NMD-targeted transcripts should increase in abundance when NMD is inhibited. We see a very strong effect for human and subtler but still significant effects for mouse, frog, zebrafish, fly, and *Arabidopsis* (Kolmogorov–Smirnov test, Bonferroni corrected $p<0.05$, Figure 3.1). The apparent magnitude of the effect of a PTC_{50nt} on NMD susceptibility varies between species and this may be due to both biological and technical reasons. As mentioned previously, the efficiency of NMD inhibition likely varies between the different species as seen by the differences in fractions of transcripts and genes that change in abundance (Table 3.3). Thus, the lower magnitude of differences for mouse,

Table 3.3 Gene and isoform expression differences upon depletion of UPF1

Species	Number of genes (CDS$^+$)			Number of isoforms (CDS$^+$)		
	FPKM>1 in either cond.	>2x increase w/ NMD inhib.	>2x decrease w/ NMD inhib.	FPKM>1 in either cond.	>2x increase w/ NMD inhib.	>2x decrease w/ NMD inhib.
Homo sapiens	11,576	1,432 (12.4%)	1,911 (16.5%)	27,391	6,922 (25.3%)	6,186 (22.6%)
Mus musculus	11,511	107 (0.9%)	145 (1.3%)	21,908	1,814 (8.3%)	1,345 (6.1%)
Xenopus tropicalis	8,929	1,665 (18.7%)	1,620 (18.1%)	13,028	3,373 (25.9%)	2,847 (21.9%)
Danio rerio	13,138	124 (0.9%)	144 (1.1%)	20,883	970 (4.6%)	926 (4.4%)
Drosophila melanogaster	7,742	337 (4.4%)	78 (1.0%)	14,620	1,675 (11.5%)	1,386 (9.5%)
Arabidopsis thaliana	15,997	635 (4.0%)	435 (2.7%)	21,930	1,801 (8.2%)	1,350 (6.2%)
Schizosaccharomyces pombe	4,698	197 (4.2%)	135 (2.9%)	5,204	298 (5.7%)	245 (4.7%)

Figure 3.1 Effect of the 50nt Rule on transcript stability



* p is lower than reportable by R

Cumulative density functions (cdfs) of $\log_2$(NMD inhibition FPKM/control FPKM) for NTC transcripts (non-PTC_{50nt}, in black) and PTC_{50nt} transcripts (in red) for each species. Kolmogorov–Smirnov tests were performed to compare the two distributions (alternative hypothesis: the CDF of NTC transcripts lies above that of PTC_{50nt} transcripts) and D statistics and corrected p-values (Bonferroni, n=7) are reported. Only transcripts with FPKM>1 in at least one of the two conditions and Cuffdiff status of 'OK' were included. The plots and tests were done in R using the plot.ecdf and ks.test functions, respectively.

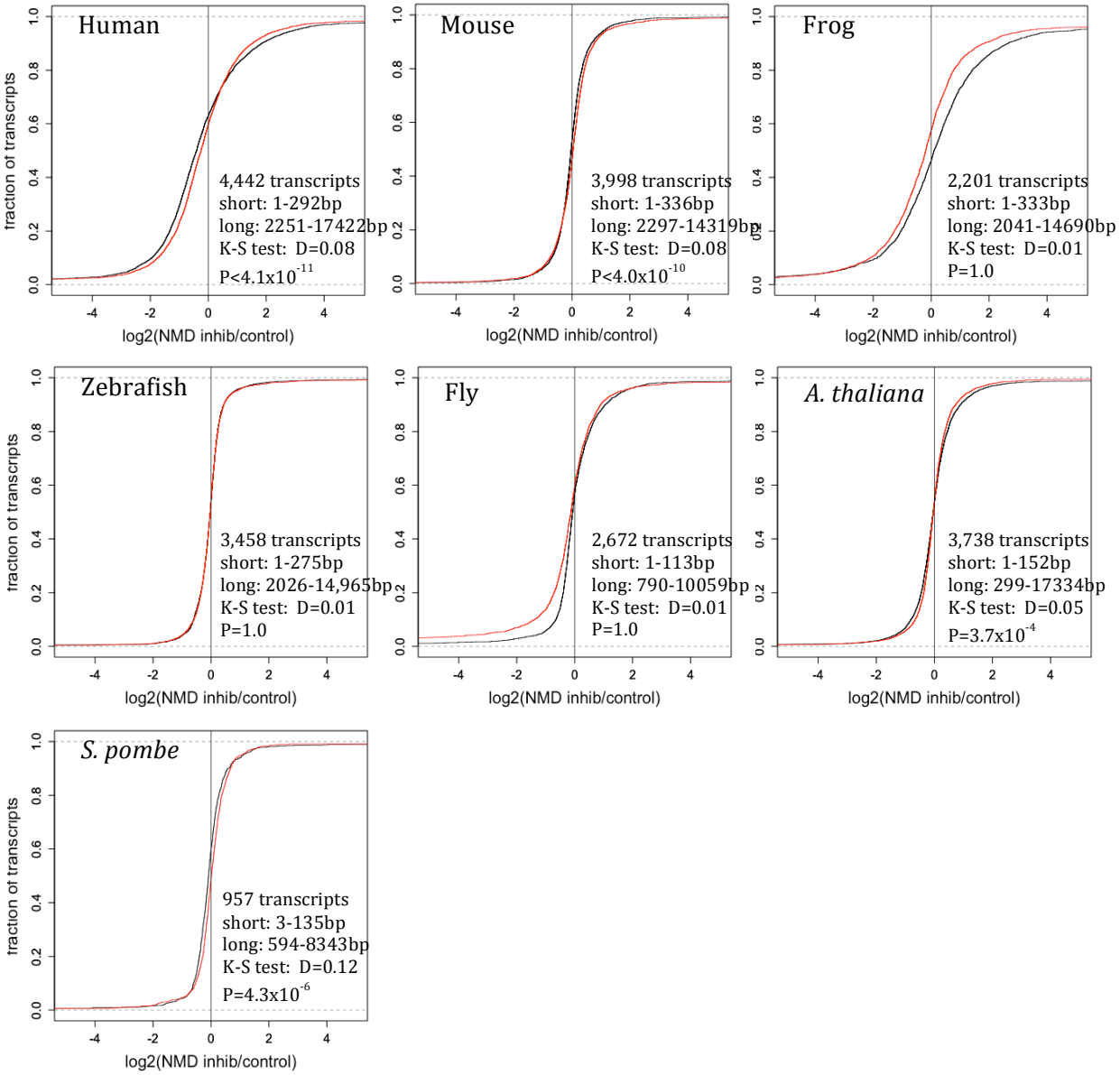
zebrafish, and *Arabidopsis* may be explained by the overall lower amount of changes in expression seen in the experiments. Additionally, this analysis relies on accurate prediction of the position of the stop codon for each transcript, which depends on the quality of the reference gene annotation used. Nonetheless, it is evident that a splice junction downstream of the termination codon plays some role in NMD in each species except for the fungal species, *S. pombe*. The majority of fungal NMD work has been performed on *S. cerevisiae* for which splicing is relatively rare and the 50nt Rule is assumed to not trigger NMD. However, *S. pombe* has substantial splicing and still no evidence of the 50nt Rule.

The other major model for defining premature termination codons is a long 3' UTR. To look at this in each species, we compared the distribution of fold change values of the set of transcripts with the shortest 20% of 3' UTRs to that of the set with the longest 20% of 3' UTRs (Figure 3.2). We filtered out the PTC_{50nt}-containing transcripts since those are substantially affected by NMD in most species and having a PTC_{50nt} strongly correlates with a longer 3' UTR. Using an analogous analysis as for comparing PTC_{50nt} transcripts to NTC transcripts, we see substantially smaller affects for long 3' UTRs, and no significant difference for frog, zebrafish, and fly. There is a significant effect for human, mouse, and *Arabidopsis* indicating that a long 3' UTR may indeed trigger NMD in these species, but the effect size is much smaller than seen for a PTC_{50nt}. 3' UTR length has a significant effect on NMD susceptibility in *S. pombe*, and since the 50nt Rule is not at play in this species, 3' UTR length may indeed be a major mechanism of NMD, despite the effect size still being small.

The other potentially NMD-inducing feature we looked at was the presence of an upstream open reading frame (uORF). uORFs that are translated are likely to induce NMD, barring downstream re-initiation of translation of the main CDS. One indication of increased likelihood of translation of a uORF is a Kozak or Kozak-like sequence around the start codon. We thus split the CDS^+ transcripts into four sets: 1) contain PTC_{50nt}, 2) contain uORF with a start codon in a Kozak-like context (strong uORF), 3) contain uORF with no Kozak-like context (weak uORF), and 4) no putative NMD inducing feature. A long 3' UTR was not considered since the effect size of this feature is very small or negligible for most species. The exception is *S. pombe*, for which transcripts in the 'no NMD feature' may be NMD targets if they have a long 3' UTR and transcripts with a PTC_{50nt} are not expected to be NMD targets.

We do in fact see enrichment for transcripts increasing in abundance compared to decreasing for the PTC_{50nt}-containing transcripts for all species, except *S. pombe* (Table 3.4). This effect is particularly strong for human and weaker for zebrafish, consistent with the observations from Figure 3.1. For transcripts with uORFs, we do see evidence that NMD may be degrading these transcripts, though to a smaller amount and not for all species. For human, the trend is as expected: strong uORF-containing transcripts have a greater bias towards increasing in abundance than weak uORF-containing transcripts, but both are more affected than those transcripts without a putative NMD-inducing feature. *S. pombe* transcripts have a similar trend, though the effect is smaller. We do see that a uORF is more likely to trigger NMD than a PTC_{50nt} in this species. For mouse, fly, and *Arabidopsis*, strong uORF-containing transcripts are affected by NMD, but weak uORFs appear to have little affect. For zebrafish, both categories of uORFs have an effect and for frog, interestingly,

Figure 3.2 Effect of 3' UTR length on transcript stability



Cumulative density functions (cdfs) of $\log_2$(NMD inhibition FPKM/control FPKM) for transcripts with short 3' UTRs (bottom 20% of 3' UTR lengths, in black) and transcripts with long 3' UTRs (top 20% of 3' UTR lengths, in red) for each species. The number of transcripts in each group and the 3' UTR size ranges are shown. Kolmogorov–Smirnov tests were performed to compare the two distributions (alternative hypothesis: the CDF of transcripts with short 3' UTRs lies above that of transcripts with long 3' UTRs) and D statistics and corrected p-values (Bonferroni, n=7) are reported. Only transcripts with FPKM>1 in at least one of the two conditions and Cuffdiff status of 'OK' were included. The plots and tests were done in R using the plot.ecdf and ks.test functions, respectively.

Table 3.4 Effect of uORFs on transcript stability

Species	NMD-inducing feature	Number of transcripts	>2x increase w/ NMD inhib.	>2x decrease w/ NMD inhib.	# increased / # decreased (normalized to non-NMD)
Homo sapiens	PTC_{50nt}	5,112	2,979	523	12.85
	Strong uORF	5,407	1,342	1,089	2.78
	Weak uORF	7,810	1,428	1,928	1.67
	None	9,062	1,173	2,646	1.00
Mus musculus	PTC_{50nt}	1,862	378	120	2.85
	Strong uORF	3,885	330	247	1.21
	Weak uORF	6,798	577	500	1.04
	None	9,363	529	478	1.00
Xenopus tropicalis	PTC_{50nt}	1,180	498	213	2.03
	Strong uORF	2,214	528	500	0.92
	Weak uORF	4,913	1,143	1,087	0.91
	None	4,721	1,204	1,047	1.00
Danio rerio	PTC_{50nt}	2,507	198	169	1.30
	Strong uORF	3,027	121	115	1.17
	Weak uORF	8,588	374	335	1.24
	None	6,761	277	307	1.00
Drosophila melanogaster	PTC_{50nt}	1,068	357	106	3.43
	Strong uORF	2,159	280	213	1.34
	Weak uORF	6,129	590	608	0.99
	None	5,264	448	456	1.00
Arabidopsis thaliana	PTC_{50nt}	2,626	483	214	1.87
	Strong uORF	725	72	48	1.24
	Weak uORF	6,686	435	417	0.86
	None	11,893	811	671	1.00
Schizosaccharomyces pombe	PTC_{50nt}	216	20	24	0.74
	Strong uORF	453	45	26	1.54
	Weak uORF	2606	131	104	1.12
	None	1,929	102	91	1.00

NMD effect on NTC-containing isoforms containing a uORF. The numbers for PTC_{50nt}-containing isoforms are included for comparison. A 'strong' uORF is one where the start codon is in a 'Kozak-like' context. The Kozak-like sequences used for the different species is as follows: vertebrates - RNNATGGV; fly - MRMMATG; *Arabidopsis* - AMNATGG; *S. pombe* - ANANAATG. A 'weak' uORF is ATG in any other context. The ratio of increased/decreased is included as a measurement of the effect on NMD on the set of isoforms. The ratio is normalized to that of the set with no obvious NMD-inducing features (no PTC_{50nt} or uORF) because any changes there are likely to be secondary and not due to the inhibition of degradation.

neither category has an effect. The results for frog are in contrast to those for all the other species in that there is no evidence that having a uORF triggers NMD in this species in this context, despite the substantial effect we see on transcripts with a PTC_{50nt}.

Pervasiveness of NMD-degraded isoforms and potential for regulation

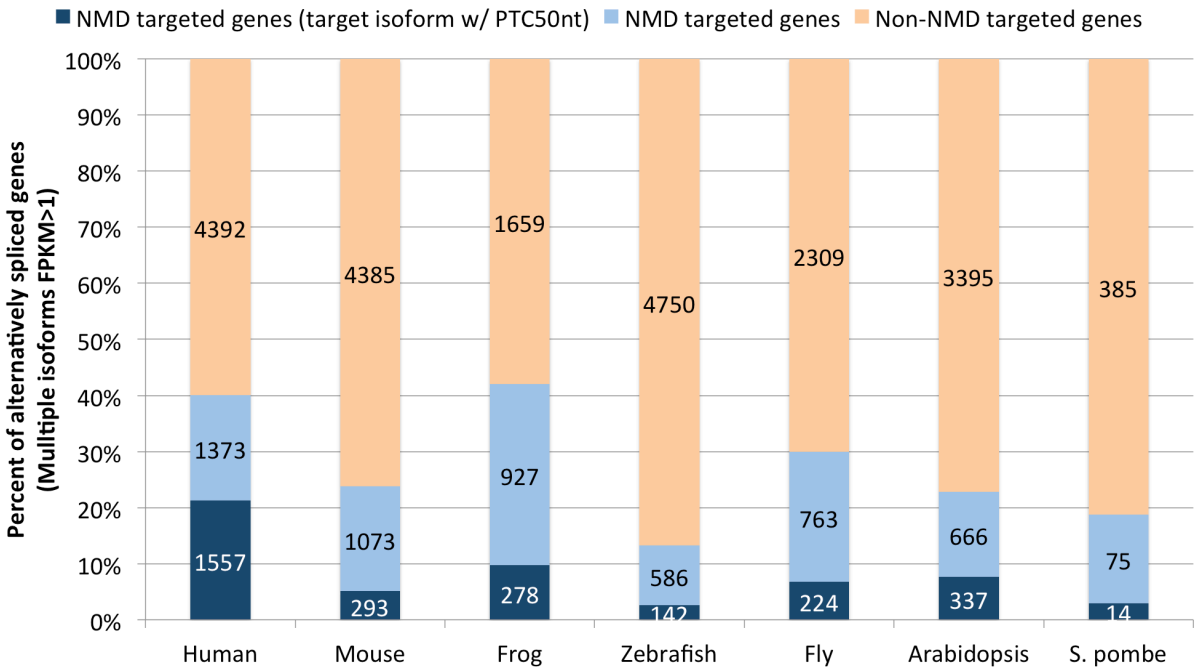
While NMD's role in gene regulation has been studied to varying extents in the species in our data set, we wanted to systematically investigate how pervasive this potential for regulation via NMD is throughout eukaryotes. We first focused on NMD-coupled to alternative splicing. To investigate this, we analyzed the genes that had more than one isoform expressed with FPKM>1 in at least one of the two conditions (control and NMD-inhibited). We found that 13-40% of these alternatively spliced genes are potential NMD targets (Figure 3.3.). An NMD-targeted gene is defined as having at least one isoform that increases at least 2-fold upon NMD inhibition (the NMD target) and at least one isoform that does not increase more than 1.2-fold (to control for transcriptional up-regulation). In fact, for most species, over 20% of alternatively spliced genes have the potential to be regulated via alternative splicing coupled with NMD. The exceptions are zebrafish (for which the NMD inhibition may not have been very efficient) and *S. pombe*. In order to find a more confident set of NMD targeted genes, we further required that at least one of the isoforms that increased in abundance contain a PTC_{50nt}, and that at least one of the non-increasing isoforms contained a NTC (Figure 3.3, dark blue). This result indicates that a minimum of 5-20% of alternatively spliced genes may be susceptible to alternative splicing coupled with NMD in each species excluding zebrafish (see above) and *S. pombe* (for which the 50nt Rule does not appear to matter).

We also looked at how NMD affects the transcriptomes of the different species independent of alternative splicing. We find that 5-40% of the genes expressed in this experiment have an isoform that increases over 2-fold when UPF1 is depleted and NMD inhibited (Figure 3.4). Since we also generally see a large number of transcripts that decrease in abundance (Table 3.3), many of these changes may be secondary effects of the depletion and not represent direct targets of NMD. A higher confidence set of direct targets is those genes with isoforms that have an NMD-inducing feature (PTC_{50nt} or uORF) and increase upon NMD inhibition. The percent of expressed genes with a PTC_{50nt}-containing isoform that increases is by far highest for human (19%), and lower for the other species (3-5% excluding zebrafish and *S. pombe*). Many more genes have a uORF-containing isoform that increases and may therefore be a direct target. Thus, while hundreds of genes in each species, except possibly *S. pombe*, produce higher confidence direct targets of NMD, the fraction of genes affected appears to vary greatly between species.

Discussion

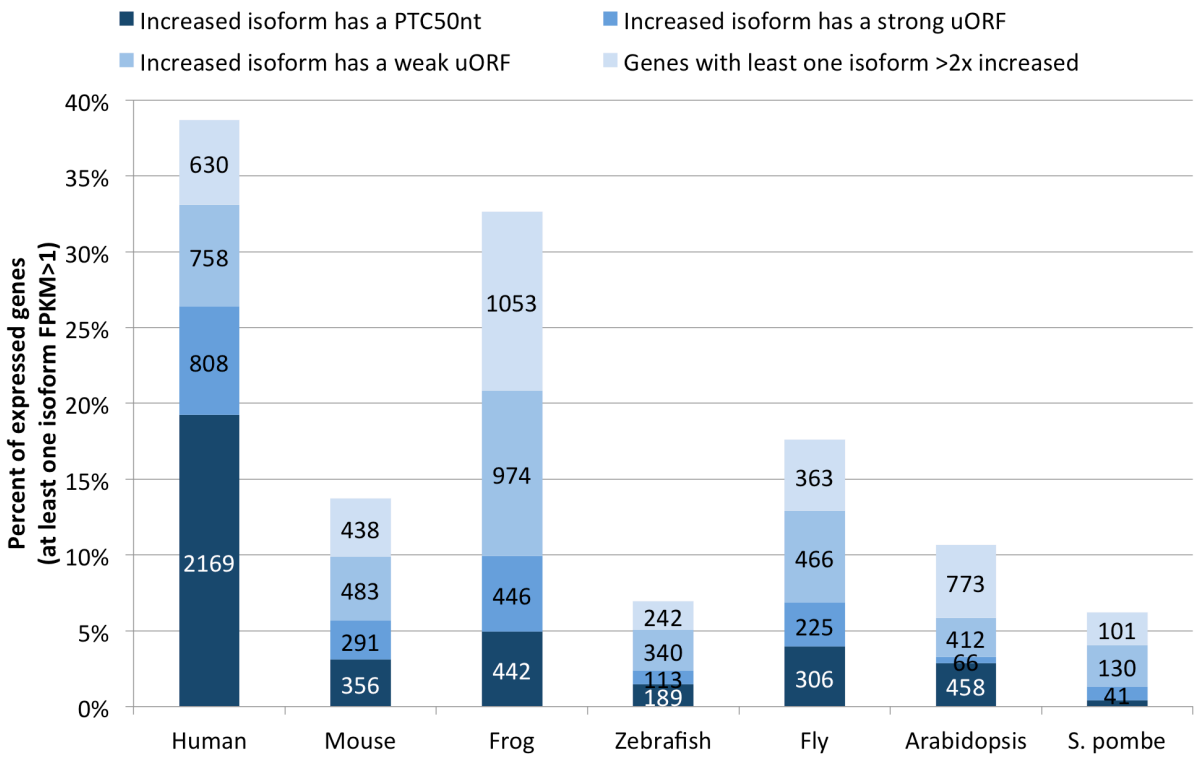
In this study, we analyzed RNA-seq data from UPF1-depletion experiments performed in seven diverse eukaryotic species including animal, plant, and fungi. We investigated the prevalence of NMD-coupled to alternative splicing and the features of NMD targets in the different species. We find that in all species studied, a substantial number of genes are targets of NMD and potentially regulated in this way. Additionally, we find evidence that an exon-exon junction >50nts downstream of a stop codon has the potential

Figure 3.3 Alternative splicing coupled to NMD in seven species



Fraction of alternatively spliced genes that are putative NMD targets for each species. Alternatively spliced genes are those that have more than one isoform with FPKM>1 in at least one of the two conditions. The percent of expressed genes (at least one isoform with FPKM>1 in at least one condition) that are alternatively spliced varies greatly between species (human: 65%, mouse: 50%, frog: 32%, zebrafish: 43%, fly: 43%, *Arabidopsis*: 27%, *S. pombe*: 10%). NMD-targeted genes (blue/dark blue) are defined as those that have at least one isoform that increases >2-fold when NMD is inhibited and at least one isoform that does not increase more than 1.2-fold. In dark blue is the subset of NMD-targeted genes for which at least one >2-fold increasing isoform has a PTC_{50nt} and at least one non-increasing isoform has a NTC. Only CDS+ transcripts with FPKM>1 in at least one of the two conditions and Cuffdiff status of 'OK' were included.

Figure 3.4 NMD-targeted genes in seven species



Fraction of expressed genes that are putative NMD targets for each species. Expressed genes are those that have at least one isoform with FPKM>1 in at least one of the two conditions. The total number of expressed genes varies greatly between species (human: 11,280; mouse: 11,415; frog: 8,929; zebrafish: 12,695; fly: 7,730; *Arabidopsis*: 15,997; *S. pombe*: 4,698). Dark blue: genes with a PTC_{50nt}-containing isoform that increases >2-fold when NMD is inhibited. Medium blue: genes with a NTC-containing isoform that increases >2-fold with a strong Kozak uORF (darker) or a weak Kozak uORF (lighter). Light blue: genes with a NTC-containing isoform that increases >2-fold when NMD is inhibited. Only CDS+ transcripts with FPKM>1 in at least one of the two conditions and Cuffdiff status of 'OK' were included.

to trigger NMD in all species except *S. pombe* and that uORFs likely play a role in NMD degradation in all the species except frog.

There are some limitations to the data sets analyzed that confound our ability to interpret differences in fractions of genes affected or effect sizes of NMD-inducing features between the different species. In general, we observe substantial variation between the species in the amount of transcriptome changes that occur upon UPF1 depletion (Table 3.3), and this is probably due to both biological and experimental reasons. Biological reasons include the fact that UPF1-depletion and/or NMD inhibition may in fact have different influences in the different species or in the different cell lines or tissues. Expression differences also capture secondary affects of depleting UPF1 (for instance, its non-NMD related functions) and inhibiting NMD (for instance, the mis-regulation of genes in other pathways), which also may vary between species. Two major experimental factors are that the efficiency of NMD inhibition may vary between experiments and that some samples were sequenced to lower read depth limiting our ability to measure expression changes. The variability in efficiency of NMD inhibition is likely to be a factor as different methods were used to deplete UPF1 and so the amount of depletion varies, and, in addition, the amount of depletion of UPF1 required to inhibit NMD to a certain extent also likely varies between species. Nonetheless, some categorical inferences may be made.

The percent of alternatively spliced genes that are most likely direct targets of NMD (increasing isoform contains a PTC_{50nt}, Figure 3.3, dark blue) is by far the highest in human (20%) compared to the other species (5-10%, excluding zebrafish and *S. pombe*). The mouse data revealed far fewer expression changes overall than human (Table 3.3) and a smaller percent of isoforms that increased in abundance had a PTC_{50nt} (21% v 43%), despite the fact both species are mammals where the 50nt Rule has been thought to be the canonical trigger of NMD. There are a few possible explanations for these observations. It is possible that NMD was less efficiently inhibited in the mouse experiment or that perhaps active NMD itself is just less efficient in mouse embryonic stem cells than in a human cell line (HeLa). There is evidence that the NMD pathway does indeed change in efficiency throughout differentiation and development[32]. Less efficient NMD inhibition and the physiological expression differences between mouse ES cells and human HeLa cells could also explain the smaller number of PTC_{50nt} transcripts we observe in the mouse data - 1,862 (8% of expressed transcripts) compared to 5,112 (18%) for human.

Another interesting comparison is between human and frog, a non-mammalian vertebrate. The frog data revealed that a similar fraction of transcripts and genes were affected by the depletion of UPF1 and NMD inhibition as in the human data (Table 3.3). It is interesting that ~40% of alternatively spliced genes have evidence of alternative splicing coupled to NMD (Figure 3.3), indicating that the pervasiveness of this mechanism has the potential to extend from human into other vertebrate species. Also of note is the smaller effect a PTC_{50nt} appears to have on NMD susceptibility in frog versus in human (Figure 3.1) and this is also evident in the fact that a smaller fraction of the potential NMD targeted genes have a PTC_{50nt} in the increasing isoform (dark blue, Figure 3.3). Despite the evidence that NMD was efficiently inhibited in the frog experiment (i.e. the relatively substantial

effect of a PTC_{50nt}), it is surprising that we do not see evidence for an NMD-inducing effect of uORFs (Table 3.4).

The results for *S. pombe* are also interesting as they vary so much from all the other species. In this species, the assumption is that NMD is very strongly or completely inhibited since the UPF1 gene is deleted and yet the effect overall on the transcriptome is relatively minor. It is possible then that NMD is not as influential in this species as the others, although the lower read depth of the samples might also be a cause. Despite this, it seems evident that having an exon-exon junction >50nts downstream of the stop codon does not appear to have an effect on NMD susceptibility in this species, while a long 3' UTR may play some role, and possibly more of a role than it does in the other species.

The analysis of these datasets revealed that for each species investigated, except yeast, the 50nt Rule has a stronger effect on NMD than 3' UTR length. The effect that NMD (coupled to alternative splicing or not) has on the transcriptomes of these diverse species likely varies, though the extent of variation is difficult to discern with these experiments. Up to 20-40% of alternatively spliced genes produce NMD targets and, at a minimum, a couple hundred genes are direct targets for each species in the data set, except for *S. pombe* for which it is dozens of genes. Further analyses and studies should be done to investigate the NMD-targeted genes in more detail across the species; specifically to look for conservation of NMD targeted genes and functional pathways.

Materials and Methods

Generation and collection of UPF1-depleted RNA-seq data sets

Human data was generated as described (see Chapter 1). Published data sets were use for mouse[25] and fly[44]. For zebrafish, UPF1 was knocked down by injecting embryos (11hr) with splice blocking morpholinos and RNA was extracted. Poly-A selected, unstranded libraries were generated on an IntegenX Apollo 324 machine and sequenced on an Illumina HiSeq 2500. For frog, UPF1 was knocked down by injecting embryos (stage 14) with splice-blocking (1 replicate) or translation-blocking (2 replicates) morpholinos and RNA was extracted. For *Arabidopsis* and *S. pombe*, RNA was extracted from UPF1 mutant samples (see Table 3.2). For frog, *Arabidopsis*, and *S. pombe*, poly-A selected, stranded libraries were made with the Illumina Tru-Seq kit and sequenced on an Illumina HiSeq 2500.

Read alignment, transcript assembly, and abundance quantification

The following genome builds and gene annotation were used for each species: human - GRCh17 and UCSC known genes (Dec 17, 2015); mouse - GRCh38 and UCSC known genes (Mar 8, 2016); frog - JGIv9.0 and XenBase annotation; zebrafish - Zv9 and Ensembl annotation; fly - dm3 and Ensembl annotation; *S. pombe* - ASM294 and Ensembl annotation; *A. thaliana* - TAIR10 and Ensembl annotation. Reads were mapped to the reference transcriptome with Bowtie v2.2.5[45] to calculate insert size mean and standard deviation for paired-end reads. The reads were then mapped to the reference genomes with Tophat v2.1.0[46] using the --GTF option, appropriate options for --library-type, --mate-inner-dist, and --mate-std-dev, and other default parameters except for *S. pombe* (--min-intron-length 20 --max-intron-length 1000). For each sample, transcript assembly was

performed using Cufflinks v2.2.1[47,48] and the appropriate options for --library-type and --GTF-guide. For each species, transcript assemblies were merged the Cufflinks sub-tool Cuffmerge. Each junction was assigned a Shannon entropy score based on offset and depth of spliced reads across all the libraries for a given species. Junctions that had an entropy score <1 and were not present in the reference annotation were changed to a reference junction if possible, else the transcript was filtered out. Transcript abundances were calculated for each sample using the Cufflinks sub-tool Cuffquant with the appropriate --library-type options and the --frag-bias-correct and --multi-read-correct parameters. Differential transcript expression between the control and UPF1-depletion samples for each species was determined with the Cufflinks sub-tool Cuffdiff and default parameters. Gene expression values were calculated by adding up the expression values reported the individual isoforms. For differential expression, a cutoff of a 2-fold difference between the conditions was used. Significance was not used because our ability to find significantly differentially expressed transcripts depends on read depth and the number of replicates, both of which vary between species and were often too low.

Defining the coding sequence of each transcript and any uORFs

Since the annotated coding sequences of transcripts may have been predicted differently between the different species, and since novel transcripts are of particular interest, the coding sequence (CDS) of each transcript was determined in the following way. For each transcript, the length of the open read frame (ORF) predicted from any overlapping annotated start codons was calculated. For each gene, the longest predicted ORF of any transcript was defined as the 'main CDS' of the gene. For each transcript for a given gene, the CDS is that which starts at the start codon of the 'main CDS'. If the transcript does not overlap that start codon, the longest ORF of the transcript is used as the CDS. For each transcript, uORFs are those that have a full ORF starting from a start codon that is upstream of the CDS.

References

1. Pulak, R. & Anderson, P. mRNA surveillance by the Caenorhabditis elegans smg genes. *Genes Dev.* (1993).
2. Isken, O. & Maquat, L. E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev.* **21,** 1833–1856 (2007).
3. Kervestin, S. & Jacobson, A. NMD: a multifaceted response to premature translational termination. *Nat. Rev. Mol. Cell Biol.* **13,** 700–712 (2012).
4. Lykke-Andersen, J. & Bennett, E. J. Protecting the proteome: eukaryotic cotranslational quality control pathways. *J. Cell Biol.* (2014).
5. Hwang, J., Sato, H., Tang, Y., Matsuda, D. & Maquat, L. E. UPF1 association with the cap-binding protein, CBP80, promotes nonsense-mediated mRNA decay at two distinct steps. *Mol. Cell* **39,** 396–409 (2010).
6. Chakrabarti, S., Bonneau, F., Schüssler, S., Eppinger, E. & Conti, E. Phospho-dependent and phospho-independent interactions of the helicase UPF1 with the NMD factors SMG5-SMG7 and SMG6. *Nucleic Acids Res.* **42,** 9447–9460 (2014).

7. Wittkopp, N. *et al.* Nonsense-mediated mRNA decay effectors are essential for zebrafish embryonic development and survival. *Mol. Cell. Biol.* **29,** 3517–3528 (2009).
8. Amrani, N., Sachs, M. S. & Jacobson, A. Early nonsense: mRNA decay solves a translational problem. *Nat. Rev. Mol. Cell Biol.* **7,** 415–425 (2006).
9. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.* **17,** 316–325 (2005).
10. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 189–192 (2003).
11. Rehwinkel, J., Letunic, I., Raes, J., Bork, P. & Izaurralde, E. Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA* **11,** 1530–1544 (2005).
12. Schweingruber, C., Soffientini, P., Ruepp, M.-D., Bachi, A. & Mühlemann, O. Identification of Interactions in the NMD Complex Using Proximity-Dependent Biotinylation (BioID). *PLoS ONE* **11,** e0150239 (2016).
13. He, F. & Jacobson, A. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annu. Rev. Genet.* (2015). doi:10.1146/annurev-genet-112414-054639
14. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* **16,** 665–677 (2015).
15. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* **23,** 198–199 (1998).
16. Zhang, J., Sun, X., Qian, Y., LaDuca, J. P. & Maquat, L. E. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol. Cell. Biol.* **18,** 5272–5283 (1998).
17. Behm-Ansmant, I., Gatfield, D., Rehwinkel, J., Hilgers, V. & Izaurralde, E. A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *EMBO J.* **26,** 1591–1601 (2007).
18. Hogg, J. R. & Goff, S. P. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell* **143,** 379–389 (2010).
19. Hansen, K. D. *et al.* Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. *PLoS Genet.* **5,** e1000525 (2009).
20. Chapin, A. *et al.* In vivo determination of direct targets of the nonsense-mediated decay pathway in Drosophila. *G3 (Bethesda)* **4,** 485–496 (2014).
21. Barberan-Soler, S., Lambert, N. J. & Zahler, A. M. Global analysis of alternative splicing uncovers developmental regulation of nonsense-mediated decay in C. elegans. *RNA* **15,** 1652–1660 (2009).
22. Peccarelli, M. & Kebaara, B. W. Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway. *Eukaryotic Cell* (2014). doi:10.1128/EC.00090-14
23. Kerényi, Z. *et al.* Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO J.* **27,** 1585–1595 (2008).

24. Shaul, O. Unique Aspects of Plant Nonsense-mediated mRNA Decay. *Trends Plant Sci.* (2015). doi:10.1016/j.tplants.2015.08.011
25. Hurt, J. A., Robertson, A. D. & Burge, C. B. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.* **23,** 1636–1650 (2013).
26. Ramani, A. K. *et al.* High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective Caenorhabditis elegans. *Genome Biol.* **10,** R101 (2009).
27. Nyikó, T., Sonkoly, B., Mérai, Z., Benkovics, A. H. & Silhavy, D. Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol. Biol.* **71,** 367–378 (2009).
28. Yepiskoposyan, H., Aeschimann, F., Nilsson, D., Okoniewski, M. & Mühlemann, O. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. *RNA* **17,** 2108–2118 (2011).
29. Lykke-Andersen, S. *et al.* Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes Dev.* **28,** 2498–2517 (2014).
30. Weischenfeldt, J. *et al.* Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biol.* **13,** R35 (2012).
31. Drechsel, G. *et al.* Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell* **25,** 3726–3742 (2013).
32. Lou, C. H. *et al.* Posttranscriptional control of the stem cell and neurogenic programs by the nonsense-mediated RNA decay pathway. *Cell Rep* **6,** 748–764 (2014).
33. Li, T. *et al.* Smg6/Est1 licenses embryonic stem cell differentiation via nonsense-mediated mRNA decay. *EMBO J.* (2015). doi:10.15252/embj.201489947
34. Anastasaki, C., Longman, D., Capper, A., Patton, E. E. & Cáceres, J. F. Dhx34 and Nbas function in the NMD pathway and are required for embryonic development in zebrafish. *Nucleic Acids Res.* **39,** 3686–3694 (2011).
35. Avery, P. *et al.* Drosophila Upf1 and Upf2 loss of function inhibits cell growth and causes animal death in a Upf3-independent manner. *RNA* **17,** 624–638 (2011).
36. Matia-González, A. M., Hasan, A., Moe, G. H., Mata, J. & Rodríguez-Gabriel, M. A. Functional characterization of Upf1 targets in Schizosaccharomyces pombe. *RNA Biol* **10,** 1057–1065 (2013).
37. Garre, E. *et al.* Nonsense-mediated mRNA decay controls the changes in yeast ribosomal protein pre-mRNAs levels upon osmotic stress. *PLoS ONE* **8,** e61240 (2013).
38. Mastrangelo, A. M., Marone, D., Laidò, G., De Leonardis, A. M. & De Vita, P. Alternative splicing: Enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci.* **185-186,** 40–49 (2012).
39. Riehs-Kearnan, N., Gloggnitzer, J., Dekrout, B., Jonak, C. & Riha, K. Aberrant growth and lethality of Arabidopsis deficient in nonsense-mediated RNA decay factors is caused by autoimmune-like response. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks195
40. Nelson, J. O., Moore, K. A., Chapin, A., Hollien, J. & Metzstein, M. M. Degradation of Gadd45 mRNA by nonsense-mediated decay is essential for viability. *Elife* **5,** (2016).
41. Zheng, S. Alternative splicing and nonsense-mediated mRNA decay enforce neural

specific gene expression. *Int. J. Dev. Neurosci.* (2016). doi:10.1016/j.ijdevneu.2016.03.003
42. Lareau, L. F. & Brenner, S. E. Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. *Mol. Biol. Evol.* **32,** 1072–1079 (2015).
43. Longman, D. *et al.* DHX34 and NBAS form part of an autoregulatory NMD circuit that regulates endogenous RNA targets in human cells, zebrafish and Caenorhabditis elegans. *Nucleic Acids Res.* **41,** 8319–8331 (2013).
44. Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res.* **21,** 193–202 (2011).
45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
46. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
47. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
48. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).

Chapter 4 Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines*

Abstract†

Variation in the expression level and activity of genes involved in drug disposition and action ("pharmacogenes") can affect drug response and toxicity, especially when in tissues of pharmacological importance. Previous studies have relied primarily on microarrays to understand gene expression differences, or have focused on a single tissue or small number of samples. The goal of this study was to use RNA-seq to determine the expression levels and alternative splicing of 389 PGRN pharmacogenes across four tissues (liver, kidney, heart and adipose) and lymphoblastoid cell lines (LCLs), which are used widely in pharmacogenomics studies. Analysis of RNA-seq data from different individuals across the 5 tissues (N = 139 samples) revealed substantial variation in both expression levels and splicing across samples and tissue types. Comparison with an independent RNA-seq dataset yielded a consistent picture. This in-depth exploration also revealed 183 splicing events in pharmacogenes that were previously not annotated. Overall, this study serves as a rich resource for the research community to inform biomarker and drug discovery and use.

Introduction‡

Variation in the expression levels and splicing of drug metabolizing enzymes, transporters, and targets, such as receptors and ion channels, has been associated with inter-individual differences in optimal drug dose, drug efficacy, and adverse drug events[1,2]. Thus, a comprehensive study of variation in the transcriptome profiles of pharmacologically relevant tissues promises to yield important insights into the molecular basis of variation in drug response. Technological advances in quantifying the transcriptome and the rapid development of high-throughput screening methodologies have led to the identification and characterization of many biomarkers of drug response[3,4]. These innovations have transformed the way we design and analyze pharmacogenomics studies and are increasingly informing development of approaches to clinical practice.

Transcriptome sequencing, or RNA-Seq, is facilitating analyses at the transcript level with an unprecedented resolution. As the technology has developed, longer reads and higher throughput have allowed for detailed evaluation of whole transcriptomes across many samples[5]. Analytical approaches have emerged, including Cufflinks[6] and DESeq[7] for gene expression analysis and DEXSeq[8], MISO[9], and JuncBASE[10] for splicing analysis. However, the use of next-generation sequencing technology for pharmacogenomics research has been limited[4,11]. Although community-wide efforts such as GTEx[12] are

---

* This chapter was co-written with Aparna Chhibber and Sook Wah Yee with contributions from Steven Brenner, Kathy Giacomini, Steven Scherer, Nancy Cox, Eric Gamazon, Marisa Medina, Deanna Kroetz and Wolfgang Sadee, and was adapted from a published work: Chhibber A*, French CE*, Yee SW*, Gamazon ER*, et al. 2016. Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *Pharmacogenomics Journal.* *co-first authors

† Co-written with Sook Wah Yee and Aparna Chhibber

‡ Primarily written by Sook Wah Yee

facilitating studies of expression quantitative trait loci (eQTLs), there has not been an application of RNA-Seq to large sample sets across diverse human tissues with a focus on genes involved in drug disposition and tissues of greater pharmacological relevance and action.

In pharmacogenomics, polymorphisms that affect expression levels or result in alternative splicing of drug metabolizing enzymes are known to have large effects on drug disposition and response. For example, UGT1A1*28 (rs8175347), with seven thymine-adenine[13] repeats in the promoter region, leads to reduced transcription rates of this enzyme and profound toxicity in patients receiving the topoisomerase inhibitor, irinotecan[14,15]. Likewise, alternative splicing of *CYP2D6* occurs frequently in human populations and is responsible for reduced activity of the enzyme[16]. Given these large and clinically important effects in drug metabolizing enzymes, a systematic study of the transcriptome with a focus on pharmacogenes is clearly needed. While several research groups have performed transcriptome profiling and alternative splicing event analyses in human cell lines and tissues[17-19], these studies are limited to single-tissue types or use pooled samples. Thus, information about inter-individual variation in gene expression and splicing from a given tissue type or inter-tissue variation is limited, despite the value of such studies in identifying biomarkers for differential drug response or toxicity.

Given these limitations, the NIH-supported Pharmacogenomics Research Network (PGRN) initiated a transcriptome sequencing project to catalog variation in gene expression and splicing across individuals in tissues and genes of pharmacologic importance. Tissues studied include liver, a key organ for drug metabolism[20, 21], kidney, the site of excretion for many drugs[22], as well as heart and adipose tissue, where pharmacogenes can affect local drug distribution and action[23]. Lymphoblastoid cell lines were also included, as they have been widely used as a cell-based model for a variety of pharmacogenomics studies[24-26]. In this article, we characterized the variability in the expression and splicing of 389 PGRN pharmacogenes across individuals and between four human tissue types and lymphoblastoid cell lines, and identified novel alternative splicing events in these samples. Further, we provide this information for community use, in the form of expression and splicing profiles for 139 individuals. This resource will be valuable for future pharmacogenomics studies as both a discovery and validation platform.

Results

The Pharmacogenomics Research Network RNA-Seq Project[*]

The Pharmacogenomics Research Network (PGRN) RNA-Seq project was designed to provide in-depth investigation of the transcriptomes of pharmacologically relevant human tissues with a focus on genes of particular interest to the pharmacogenomics community (Figure 4.1). In order to study inter-individual variability in expression and splicing of PGRN pharmacogenes, we generated transcriptome data from 24 liver, 20 kidney, 25 heart, and 25 adipose samples and 45 LCLs. For each sample, reads were mapped to the human genome[35], resulting in 10 to 97 million mapped reads per sample. To

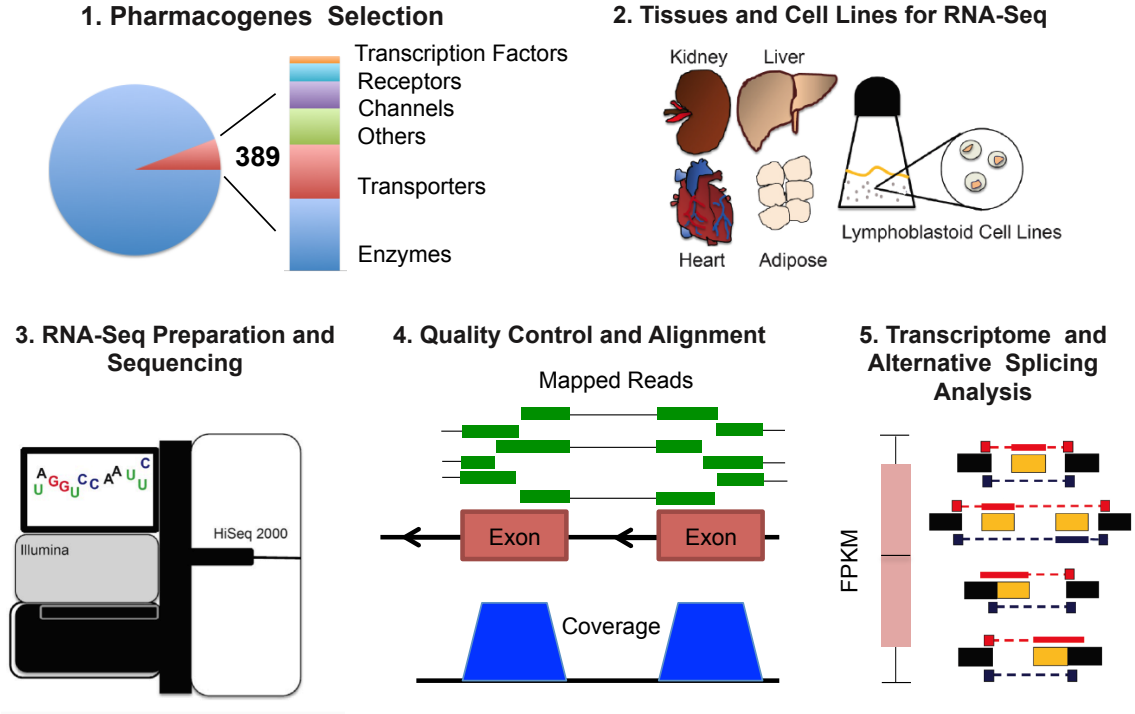
[*] Primarily written by Sook Wah Yee

control for this substantial difference in sequencing depth and sample number between tissues, 18 samples for each tissue were selected and subsampled down to 20 million reads/sample for further expression and splicing analyses, resulting in a total of 90 samples. Gene expression and splicing results are available for download for all samples at http://pharmacogenetics.ucsf.edu/expression/rnaseqdata.html and at doi:10.6078/D1RG66.

Analysis of PGRN pharmacogene gene expression[*]

We found that 161 (of 389) of our PGRN pharmacogenes were expressed at FPKM ≥1 in at least one sample across all five tissue types in our data set and 87 pharmacogenes were expressed at FPKM ≥1 in all samples of all five tissue types (Table 4.1). As a group,

Figure 4.1 Overview of the Pharmacogenomics Research Network RNA-seq project



1. 389 "PGRN pharmacogenes" were selected representing genes that play a key role in drug disposition. 2. RNA from multiple samples for human liver, heart, kidney, adipose tissue, and lymphoblastoid cell lines was collected. 3. cDNA libraries were prepared from these samples and sequenced using an Illumina HiSeq 2000. 4. Rigorous pre- and post- alignment quality control procedures were applied to the data. 5. Gene expression was quantified and splicing events identified for the PGRN pharmacogenes across samples and tissue types. This information is provided as a resource to the pharmacogenomics community. Figure by Sook Wah Yee.

[*] Analysis and writing by Aparna Chhibber

Table 4.1 Summary of gene expression for the tissues

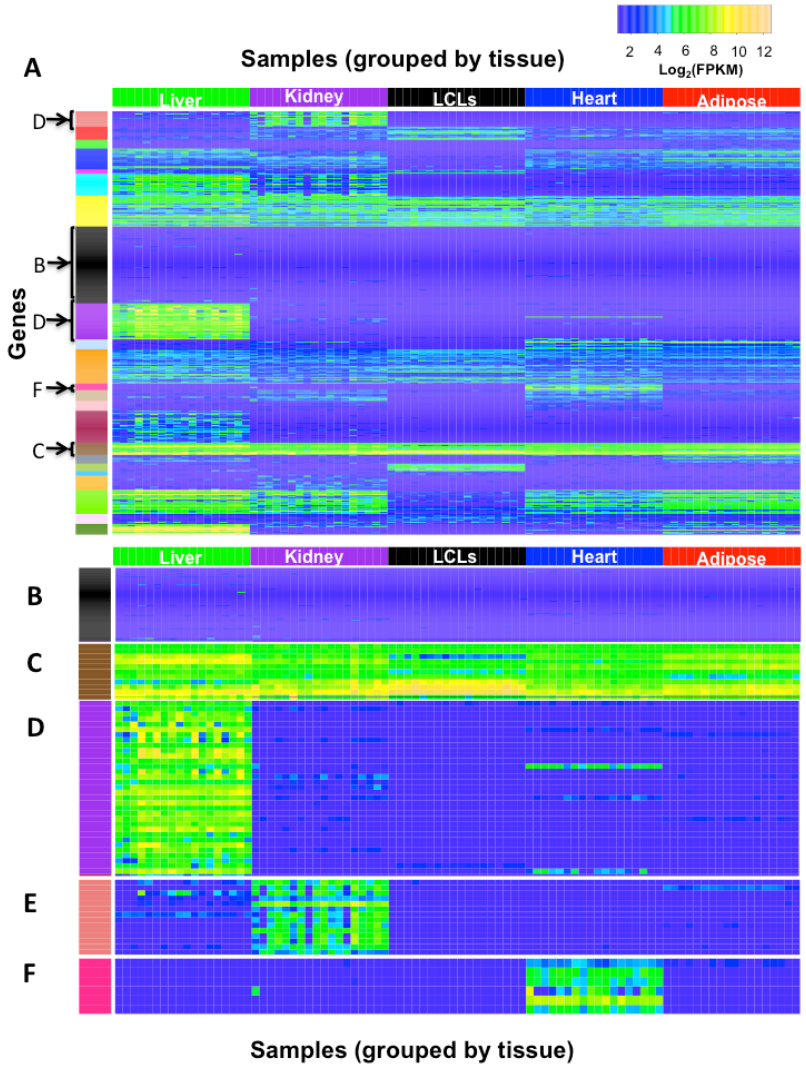
PHARMACOGENES	LCL	Liver	Kidney	Adipose	Heart	Intersection	Union
Total Number of Genes Mapped and Analyzed	389	387*	389	389	389	389	389
Ubiquitous[1]	116	225	167	190	166	87	291
Total Expressed[2]	188	320	315	274	255	161	364
Total Undetected[3]	17	3	5	1	5	0	22
Number of Specific Genes[3]	3	39	14	3	9	NA	NA
% Ubiquitous (Ubiquitous/Total)	29.80%	58.10%	42.90%	48.80%	42.70%	22.40%	74.81%
% Ubiquitous (Ubiquitous/Total Expressed)	61.70%	70.30%	53.00%	69.30%	65.10%	54.00%	79.95%
% Specific (Specific/Total Expressed)	2.10%	14.00%	5.30%	1.30%	4.00%	NA	NA

* Fewer number of genes mapped and analyzed in liver samples, because two pharmacogenes, ALB (albumin) and SERPINA1 (serpin peptidase inhibitor, clade A) were not able to be accurately quantified due to their high expression, and so they were excluded. [1] Genes that have FPKM values ≥ 1 in all 18 individuals of each tissues or LCL. [2] Total Expressed genes refers to genes that have FPKM ≥ 1 in any 18 individuals of each tissues or LCL. [3] Total undetected genes refers to genes that have FPKM = 0 in all 18 individuals of each tissues or LCL. [4] Number of specific genes refers to genes that have FPKM values ≥ 1 in one tissue only and other tissues have FPKM values < 1.

PGRN pharmacogenes were significantly enriched for variable gene expression between individuals, and were among the top ten most variably expressed gene sets (classified by gene ontology biological process[37]) in the physiological tissues. We also observed subsets of pharmacogenes that showed similar patterns of expression across the different tissues (k-means clustering of gene expression of 389 pharmacogenes, Figure 4.2A). For example, some pharmacogenes were expressed consistently at low levels across all tissues and samples (e.g. ABCC12 and ESR2, Figure 4.2B). In contrast, 11 pharmacogenes were very highly expressed, though to different levels, across all tissues and LCLs (Figure 4.2C); these include genes involved in mitochondrial structure or function (*ADH5*, *ALDH2*, *CYB5R3*, *SOD2*) and glutathione transferase activity (*GSTK1*, *GSTO1, GSTP1*)[37].

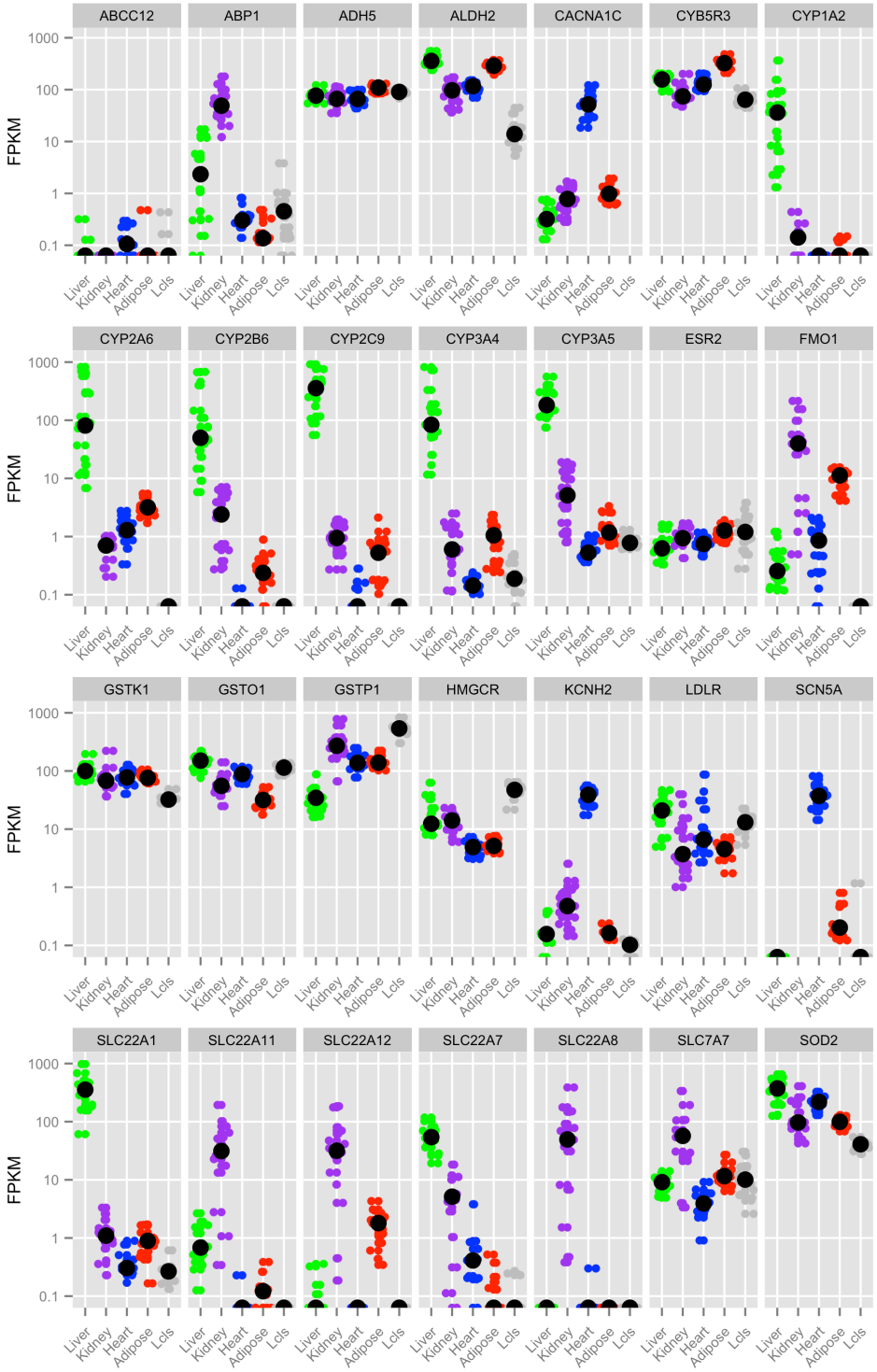
Not surprisingly, PGRN pharmacogenes are generally more highly expressed in liver compared to the other tissues. Many genes coding for xenobiotic metabolizing enzymes and transporters were highly and specifically expressed in the liver, an organ important for drug metabolism (Figure 4.2D). Pharmacogenes expressed at highest abundance in the

Figure 4.2 Heatmap of the 389 PGRN pharmacogenes' expression across 90 samples



(A) Heatmap of the 389 PGRN pharmacogenes' expression (FPKM) across 90 samples. Samples are arranged horizontally, grouped by tissue. Pharmacogenes are arranged vertically, grouped by clusters identified by *K*-means clustering; clusters are indicated by colors along the left side of the heatmap. Selected clusters show (B) genes expressed at low levels across all samples (*ABCB5, ABCC12, ABCC8, ADH7, ADRB3, ALDH3A1, BDNF, CACNA1S, CFTR, CHRM3, CHST13, CHST4, CHST5, CHST6, CHST8, CRHR1, CYP11B1, CYP11B2, CYP26A1, CYP26C1, CYP2A13, CYP2F1, CYP2S1, CYP4F8, CYP4Z1, CYP7A1, DRD1, DRD2, DRD3, DRD4, DRD5, ESR2, FMO6P, GNB3, GRM3, GSTA3, GSTA5, GSTT2, HTR1A, HTR2A, IL28B, KCNE2, MMP3, OPRM1, P2RY1, PNMT, PRSS53, RYR1, SCN3B, SLC10A2, SLC22A13, SLC22A14, SLC22A16, SLC22A4, SLC28A2, SLC28A3, SLC6A3, SLC6A4, SLCO1A2, SLCO6A1, SULT1A3, SULT4A1, TPH1, TPH2, TPSG1, UGT1A10, UGT1A5, UGT1A8, UGT2B11, UGT2B28)* (C) genes highly expressed across all samples (*ADD1, ADH5, ALDH2, CYB5A, CYB5R3, GSTK1, GSTO1, GSTP1, HLA-B, RPL13, SOD2*) or genes expressed at higher levels in (D) liver (*ABCB4, ABCC2, ADH1A, ADH4, APOA4, APOB, CYP2A6, CYP2B6, CYP2C18, CYP2C8, CYP2C9, CYP2D6, CYP2J2, CYP3A4, CYP3A5, CYP4F11, CYP8B1, F2, F5, MAT1A, NAT2, PON1, PON3, SERPINA7, SLC22A1, SLCO1B1, SLCO1B3, SULT2A1, UGT1A1, UGT1A4, UGT2B10, UGT2B15, UGT2B4)*, (E) kidney (*ABP1, FMO1, GSTA2, GSTO2, HSD11B2, SLC13A1, SLC13A3, SLC22A11, SLC22A12, SLC22A2, SLC22A6, SLC22A8, SULT1C2, UGT8*), or (F) heart (*ADRB1, CACNA1C, KCNH2, NPPB, RYR2, SCN5A*). Gene names are listed in order from top to bottom in each cluster in the figure. Plot drawn using R package gplots.[74] Figure by Aparna Chhibber.

Figure 4.3 Gene expression by sample across each tissue for selected genes



Gene expression (FPKM) by sample across each tissue type and LCLs for selected cytochrome P450 (CYP) enzymes, solute carrier family (SLC) transporters, and other pharmacogenes discussed in this article from subsampled data (18 samples per tissue type, 20 million reads per sample). The black dot indicates median FPKM per gene and tissue type. Plots drawn using R package ggplot2.[75] Figure by Aparna Chhibber.

kidney, the major organ for secretion and reabsorption, include a number of solute carrier transporters (SLC genes) (Figure 4.2E), which play important roles in drug secretion or reabsorption[38], as well as enzymes such as *ABP1* and *FMO1*. In addition, pharmacogene expression levels in the liver and kidney varied greatly among individuals. For example, the expression levels of a number of CYPs in the liver and SLC transporters in the kidney varied by over 100-fold and 1000-fold respectively (Figure 4.3).

The list of PGRN pharmacogenes included 119 (out of 389) genes that are currently drug targets or are under clinical development as potential targets for various diseases[39]. These drug target genes may be expressed abundantly in tissues not primarily involved in drug disposition. For example, a small number of pharmacogenes were highly expressed solely in the heart (Figure 4.2F). These genes are all involved with cardiac contractility and include, for example, ion channels involved in cardiac conductance (*SCN5A*, *CACNA1C*, *KCNH2*) that are targeted by many drugs[40-42]. Most pharmacogenes expressed (FPKM ≥1) in adipose tissue were expressed in other tissues as well (Figure 4.2A). The strongest correlation of pharmacogene expression profiles among tissues were detected between adipose and heart (r=0.83), as is true for all protein-coding genes expression between adipose and heart (r=0.9, Table 4.2). Compared to the four physiological tissues, LCLs showed lower overall expression levels of pharmacogenes: proportionally fewer pharmacogenes were expressed in at least one LCL sample or expressed in all LCL samples compared to all protein-coding genes (chi-square test: 48% vs. 64%, $p<0.0001$ and 30% vs. 48%, $p<0.0001$ in at least one sample or all samples respectively, Table 4.1). Pharmacogenes expressed at lower levels in LCLs than in the tissues assayed include genes important for drug disposition - for example, genes coding for enzymes (cytochrome P450s, UGTs, SULTs), SLC transporters, ion channels and receptors.

Analysis of PGRN pharmacogene splicing

We found that 278 of the 389 pharmacogenes (72%) showed clear evidence of being alternatively spliced (≥ 2 isoforms) in our data set. Receptor and channel genes are the least alternatively spliced (<50%, Table 4.3), although, likely due to the small numbers of genes, only receptors are significantly depleted (Bonferroni-corrected $p<0.05$, hypergeometric test). Another 66 pharmacogenes had inconclusive evidence of being alternatively spliced either because the alternative splice event is very rare, or because of low gene expression. The other 45 pharmacogenes are substantially expressed (FPKM>10) in at least one sample but have no evidence of alternative splice events in this dataset.

Differential alternative splicing between pairs of tissues was evident for dozens of PGRN pharmacogenes (Wilcoxon test, FDR<0.05, difference in median PSI >5) (Table 4.4, with LCLs showing the greatest differences in splicing events compared with the other tissues. We also found dozens of inferred splice events that were only observed in one of our five tissue types, often because the gene was not expressed in other tissues but also possibly because only alternative splice events were used in those tissue types (Figure 4.4A). When we control for gene expression differences between tissues by requiring the potentially alternatively spliced region to have high total read coverage in a number of samples for the four other tissues, we see only a very small fraction of genes (0-5%) have tissue-specific splice events.

Table 4.2 Differential gene expression between tissues

	Liver	Kidney	Adipose	Heart	LCL
Liver		**0.40** (U:119/D:69)	**0.43** (U:134/D:48)	**0.31** (U:141/D:58)	**0.27** (U:173/D:38)
Kidney	**0.87** (U:3540/D:3023)		**0.64** (U:94/D:51)	**0.61** (U:94/D:48)	**0.51** (U:151/D:39)
Adipose	**0.85** (U:3707/D:3862)	**0.87** (U:2940/D:3576)		**0.83** (U:59/D:46)	**0.62** (U:117/D:38)
Heart	**0.82** (U:3828/D:3802)	**0.85** (U:2966/D:3335)	**0.9** (U:3118/D:3275)		**0.62** (U:115/D:39)
LCL	**0.76** (U:5113/D:3713)	**0.76** (U:5198/D:4265)	**0.77** (U:5848/D:3781)	**0.75** (U:5360/D:3907)	

A total of 20 pairwise comparisons were performed to detect the genes differentially expressed between LCLs and tissues for all protein coding genes (red) and pharmacogenes (blue). The **bold** numbers are the gene expression (FPKM) spearman correlations between two tissues for all genes. The numbers inside the parentheses refer to the number of genes differentially expressed between pairs of tissues (U: higher in row tissue/D: higher in column tissue). Genes are considered to be differentially expressed when $q<0.10$ in both DESeq and Cuffdiff and >2 fold difference in expression. Darker shading indicates higher correlation.

Table 4.3 Alternatively spliced pharmacogenes by class

class	#genes	alt. spliced genes	not alt. spliced genes	% alt. spliced
Other_Transporter	4	4	0	100.0%
ADH_Metabolism	7	6	0	85.7%
Other_Metabolism	54	45	5	83.3%
ABC_Transporter	24	20	1	83.3%
UGT_Metabolism	18	15	1	83.3%
SULT_Metabolism	10	8	0	80.0%
SLC_Transporter	56	43	5	76.8%
Other	77	55	9	71.4%
ALDH_Metabolism	7	5	0	71.4%
GST_Metabolism	17	12	1	70.6%
Nuclear Receptor/TF	24	16	7	66.7%
CYP_Metabolism	47	30	6	63.8%
Receptor	29	13	5	44.8%
Channel	15	6	5	40.0%
Total	**389**	**278**	**45**	**71.5%**

Percent of PGRN pharmacogenes that are alternatively spliced by class. Alternatively spliced genes have multiple mutually exclusive junctions with at least 1 read/100bp and PSI>5 in at least of the 90 samples. Genes reported as not alternatively spliced were not reported to have any evidence of mutually exclusive junctions in any sample, and have gene FPKM>10 in at least one sample. For other genes, it's unclear because the PSI may be below 5 or the gene may be low expressed and so some alternative splice events may not be observed.

Notably, a total of 183 alternative splicing events (in 102 out of 389 genes) included splice junctions not previously annotated, but which were present with a robust coverage of at least 5 reads/100bp in at least one sample (Figure 4.4B). The greatest number of previously non-annotated pharmacogene splicing events was observed in the liver samples, likely because many of those genes are very highly expressed in that tissue, making it easier to observe these often low expressed events. One of the novel splicing events observed in liver was an alternative last exon of *SLC22A7*, a gene that encodes a transporter of endogenous compounds and prescription drugs (Figure 4.4C). This newly found alternative event was validated by PCR*, and is predicted to produce a protein with a

* Experiment performed by Sook Wah Yee

Table 4.4 Differential splicing between tissues

	Liver	Kidney	Adipose	Heart	LCL
Liver		**18** (6.3%)	**14** (8.2%)	**10** (8.8%)	**14** (15%)
Kidney	**590** (9.2%)		**12** (10%)	**8** (6.1%)	**14** (16%)
Adipose	**849** (14%)	**826** (12%)		**12** (13%)	**19** (23%)
Heart	**787** (15%)	**743** (12%)	**1021** (16%)		**16** (21%)
LCL	**1246** (21%)	**1408** (19%)	**1538** (21%)	**1403** (21%)	

A total of 20 pairwise comparisons were performed to detect the genes differentially alternatively spliced between LCLs and tissues for all protein coding genes (red) and pharmacogenes (blue). Reported are the number of genes (**bold**) with a significantly differentially spliced event (in parentheses is the percent of tested events that were significant). Significant splicing events were determined by a Wilcoxon test on the 'percent spliced in' (PSI) values between two tissues (Benjamini-Hochberg corrected p-value <0.05). To be tested, an event must have coverage of at least 10 reads/100bp in at least half the samples for each of the two tissues, and a delta PSI between the tissues of at least 5. Darker shading indicates a smaller percentage of events are differentially spliced.

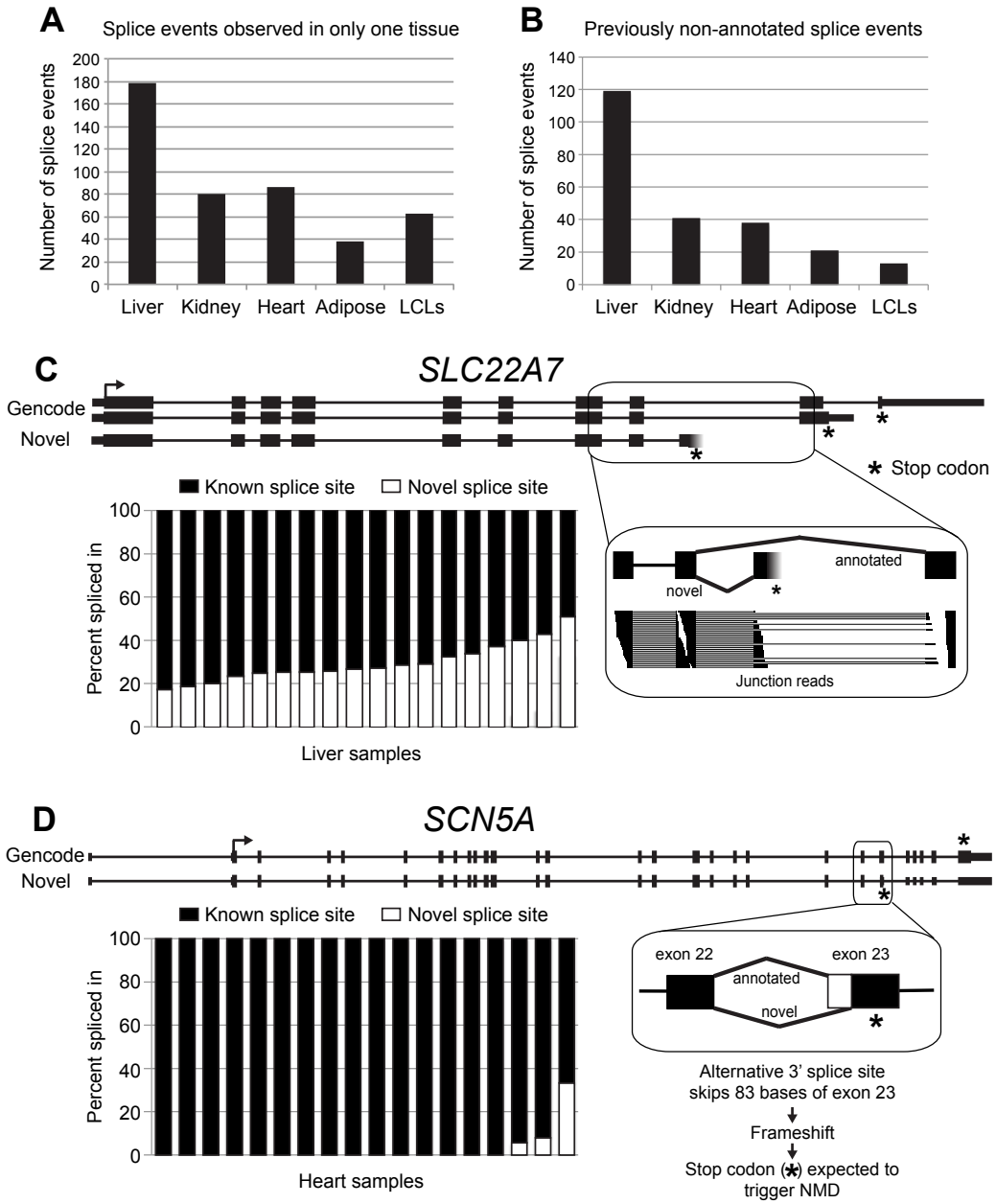
truncated C-terminus, and was substantially and variably expressed in the liver samples. A novel splicing event observed in heart was in *SCN5A*, a gene encoding a sodium channel important in maintaining normal cardiac rhythm (Figure 4.4D). Observed in three heart samples, this novel alternative 3' splice site in exon 23 excludes 83 bases and generates a downstream premature termination codon that is expected to cause the transcript to be degraded by the nonsense-mediated mRNA decay pathway.

Discussion*

Over the last several years, there have been many studies using RNA-Seq to quantify gene expression and to identify novel alternative splicing events in many tissue and cell types[43-48]. Here, we applied this approach to characterize the expression of 389 genes of pharmacologic importance (genes involved in drug disposition, response or toxicity) in multiple human tissue types and LCLs. Unlike many other transcriptome profiling studies using RNA-Seq, this report presents findings for multiple samples across tissues, allowing the capture of inter-individual variation in expression levels in addition to comparison of expression and splicing across different tissues. Further, results in multiple subjects act as biological replicates for a given tissue type, allowing for a more accurate variation in our representation of tissue-specific splicing and expression. By incorporating

* Co-written with Aparna Chhibber and Sook Wah Yee

Figure 4.4 Tissue-specific and/or novel splice events in pharmacogenes



(**A**) Splice events in PGRN pharmacogenes with PSI (percent spliced in) ≥5 and coverage ≥1 reads/100bp in at least one sample of one tissue and no coverage in any of the four other tissues. (**B**) Splice events in pharmacogenes not present in current gene annotations with coverage ≥5 reads/100bp in at least one sample. These splice events were identified in 68, 31, 18, 16, and 10 pharmacogenes in liver, kidney, heart, adipose tissue and LCLs respectively. (**C**) An alternative last exon in *SLC22A7*, not previously annotated, was observed in liver samples and would alter the C-terminal end of the protein. Chart: fraction of transcripts from *SLC22A7* that contain the novel (white) or known (black) splice event in each liver sample. Inset: reads crossing the alternative junctions in a liver sample. (**D**) A novel alternative 3' splice site in *SCN5A* was identified that results in an 83-base deletion of the coding sequence of *SCN5A*, creating a premature stop codon expected to trigger nonsense-mediated mRNA decay. Chart: fraction of transcripts from *SCN5A* that contain the novel (white) or known (black) splice event in each heart sample.

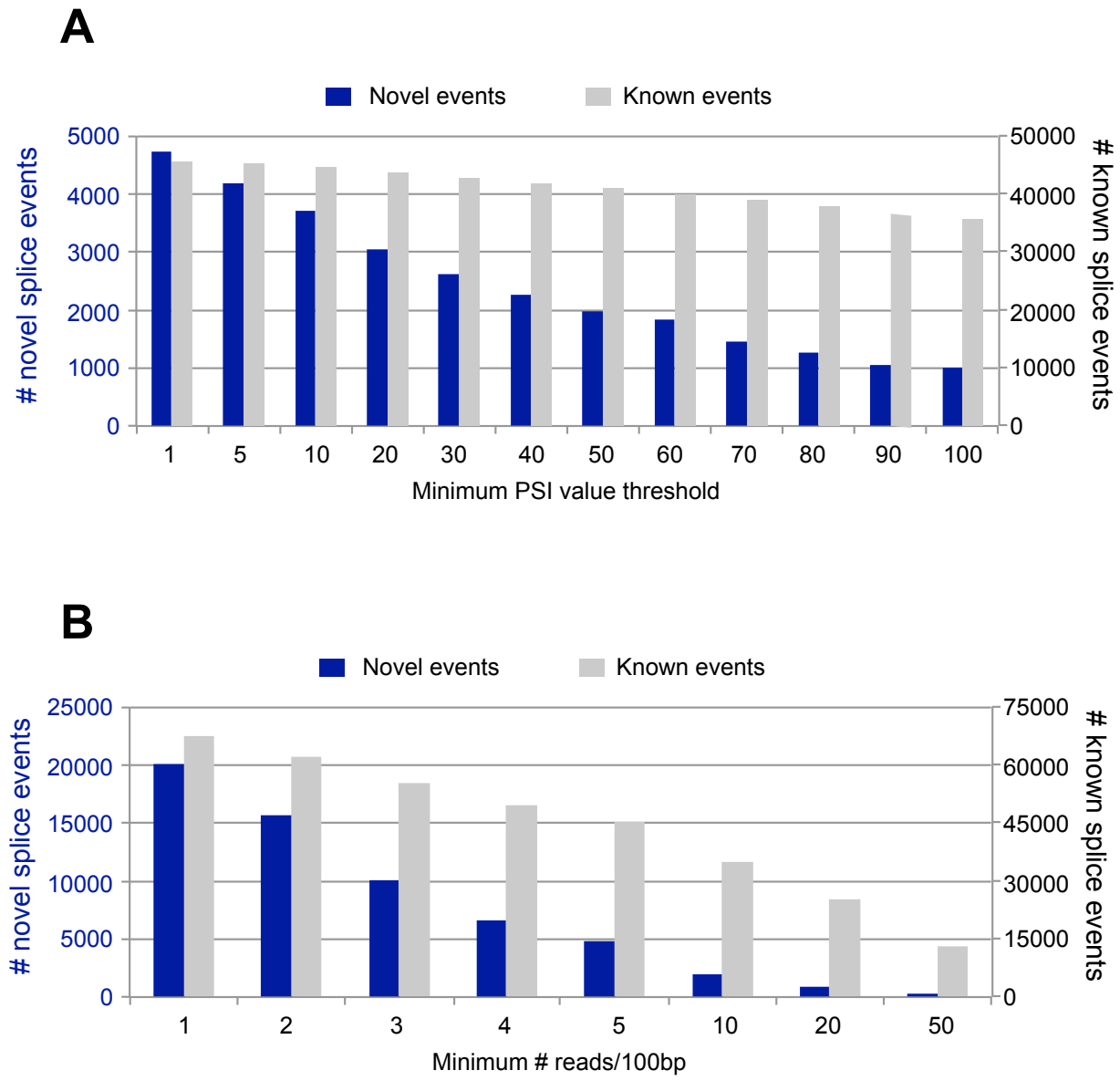
inter-individual variation in our study of several human tissues, our data represent an important addition to our understanding of human transcriptomics. This dataset is available at http:// pharmacogenetics.ucsf.edu/expression/rnaseqdata.html (and at doi:10.6078/D1RG66).

In comparing global analyses of protein-coding and pharmacogene expression, we observed several interesting patterns. Prominently, the majority of PGRN pharmacogenes were expressed at lower levels in LCLs compared to the four physiological tissues studied, in contrast to expression levels across all protein coding genes. As an actively and aggressively proliferating cell type, gene expression in LCLs is tuned to growth, and thus relative expression of genes involved in other cellular processes may be suppressed. Further, it is possible that peripheral B-lymphocytes, the primary cells from which LCLs are derived, also show significantly different patterns of expression from the other four physiological tissues included in this study. These results suggest that consideration of the phenotype and gene of interest is important when using LCLs as a proxy for other tissues in pharmacogenetic studies, as well as when using tissues as proxies for each other. Overall, more pharmacogenes were expressed at higher levels in the liver compared to other tissues. While this result is not unexpected given the importance of the liver in drug metabolism and transport and the bias towards liver-specific genes in the field of pharmacogenomics, it also demonstrates the importance of conducting studies in samples of the relevant tissue type where possible. We also observed high correlation in gene expression values between adipose and heart tissues among both protein coding genes and the subset of PGRN pharmacogenes. This result is consistent with the finding that adipose derived stem cells have been shown to spontaneously differentiate into cardiomyocytes and that both adipose and cardiac tissues derive from the mesoderm[49, 50].

We also observed interesting patterns of alternative splicing in this study, including the discovery of splicing events not previously annotated and significant differential splicing between LCLs and other tissue types. Since splicing detection is dependent on sequencing coverage and the number of samples analyzed, we investigated the effects of subsampling down the number of reads and samples to make them equivalent between tissues (Figure 1.5). Using only 18 samples per tissue, we were still able to detect 95% of splice events we would be able to observe with all samples in our data set. Subsampling the reads limits the detection of rare splicing events which particularly affects novel splice events, as they generally have low PSI values and, thus, low read coverage (Figure 4.5), or occur in only a small number of samples. As rare splice events may represent physiologically relevant alternative splicing, the splicing results from using all of our data (all reads, all samples) are also available to download.

Among the splicing events identified, we observe both previously characterized as well as novel alternative splicing events. For example, an alternative 3' splice site in the drug target *SCN5A* generates a premature termination codon predicted to trigger the nonsense-mediated mRNA decay pathway (NMD). *SCN5A* encodes the main cardiac voltage-gated sodium channel important in maintaining normal cardiac conduction. A number of drugs target sodium channels, including antiarrhythmics and non-

Figure 4.5 Novel junctions have lower PSI values and read coverage



(**A**) The distribution of the number of previously non-annotated splice events observed in at least one sample (of 90 samples) with PSI greater than the threshold. All events also have read coverage of at least 5 reads/100bp. (**B**) The distribution of the number of non-annotated splice events observed in at least one sample (of 90 samples) with read coverage greater than the threshold.

antiarrhythmic sodium channel blockers. Changes in structure, activity, and expression of drug targets, such as that encoded by SCN5A, can alter the efficacy of drugs designed to target these proteins[51, 52]. This event may be indicative of a novel role for alternative splicing coupled with NMD in the regulation of this gene[53]. Additionally, a novel truncated isoform of the transporter *SLC22A7* was identified. The gene *SLC22A7* is expressed in both kidney and liver and is important for transport of endogenous compounds [54] and a number of prescription drugs[55-57].

We also observed substantial variability in gene expression, particularly among drug transporters and drug metabolizing enzymes. In the liver, several cytochrome P450 (CYP) enzymes showed significant variability in expression levels between individuals; such variability can drive differences in drug metabolism across individuals, leading to variation in drug efficacy and susceptibility to toxicity[58]. One example includes CYP3A4, which is responsible for activation and deactivation of a number of drugs by oxidation in the liver. Induction of CYP3A4 by concomitant medications or dietary supplements is well-established, and is considered a major source of variation in drug response[59]. The enormous inter-individual variation in the expression levels of CYP3A4 we observe in the liver samples may be due to differences in diet, including dietary supplements, or medications among the individuals, in addition to genetic variation. Like drug metabolism, renal elimination of drugs is also variable across individuals in part due to variation in renal secretion and reabsorption; this variation can be driven by differences in expression levels of renal transporters across individuals. We observed profound differences in the expression levels of renal secretory and reabsorptive transporters, particularly the solute carrier transporters (SLCs). For example, expression of the uric acid transporter *SLC22A12* varied almost 1000-fold between individuals in the kidney (Figure 4.3). As a target for drugs that treat hyperuricemia[60], the expression level of *SLC22A12* could be an important determinant of drug response.

As is true of any study using human organs, while only healthy tissues were used for mRNA extraction, the patients themselves may have had a disease affecting other organs or may have been taking medications. In particular for the study of pharmacogenes, the variability in xenobiotic exposure is a concern, as such exposure is known to alter pharmacogene expression[61, 62] and splicing[63-67] profiles. The fact that the variability in splicing and expression both within and between tissues was similar to that identified in an analogous analysis of an independently derived RNA-seq dataset (from the GTEx[36] project) suggests that the patterns of splicing and expression detected are not driven by a single overrepresented disease, phenotype, or environmental exposure in our dataset. However, in both datasets the variability detected may be driven in part by differences in health status or exposures between individuals. Other potential sources of variability in our dataset include subtle differences in cellular composition of the tissue samples or sample collection protocols, as well as patient age and sex[68, 69]; for example, a few pharmacogenes appeared to show higher expression levels in samples from pediatric patients. Given the small sample sizes and skewed sex and age distributions in some of the tissue types, this study was not optimal for investigating variation due to these two factors. Finally, despite substantial variability in expression in some pharmacogenes between individuals, other pharmacogenes showed very consistent expression between tissues and/or across

individuals (e.g. ADH5 and GSTK1), suggesting that the extensive variability observed was not driven by noise in the experimental process.

Pharmacogenomic studies have largely focused on the effects of genetic polymorphisms in pharmacogenes on drug response and drug toxicity[70, 71]. Our data suggest that genes involved in drug disposition and toxicity can be variably spliced and expressed among individuals and across tissues. Further, given that splicing can affect expression, localization, and function of genes[72, 73], our results suggest that splicing may be a relatively unexplored source of variability in drug response, toxicity, and efficacy. Transcriptome profiling (including both expression and splicing) of pharmacogenes may be a valuable tool for identification of mechanisms and possible prediction of this variability. As the first in depth analysis of transcript structure and expression of genes that play a key role in drug disposition, this PGRN RNAseq resource will be valuable for biomarker and drug target discovery and validation.

Materials and Methods*

Selection of Pharmacogenes

Protein coding genes were defined as those with a start codon in the Gencode v12[13] annotation. A subset of these was defined as "PGRN pharmacogenes." Our list of 389 pharmacogenes was compiled from PharmGKB[27], a curated knowledgebase about the impact of genetic variation on drug response, PharmaADME[28], the U.S. Food and Drug Administration (FDA) Pharmacogenomics Biomarkers[29] and the literature[24, 30-33]. Genes that are annotated in at least two of these resources or publications were selected as PGRN pharmacogenes. These include 160 enzymes, 84 transporters, 15 ion-channels, 27 receptors, 24 nuclear receptors and other transcription factors, as well as 22 other genes, including G-protein coupled receptors that are drug targets and play an important role in drug disposition, response, or toxicity.

Tissues Collection, RNA Isolation, and Preparation of RNA-sequencing Library

Tissue from 24 liver, 20 kidney (cortex), 25 heart (left ventricle), 25 adipose (subcutaneous) samples, and 45 lymphoblastoid cell lines (LCLs) were obtained from PGRN research groups: the Pharmacogenomics of Anticancer Agents Research in Children (PAAR4Kids) provided liver tissues, Pharmacogenomics of Membrane Transporters (PMT) provided kidney samples, Pharmacogenomics and Risk of Cardiovascular Disease (PARC) provided adipose tissue and lymphoblastoid cell lines, and Pharmacogenomics of Arrhythmia Therapy (PAT) provided heart tissue. Total RNA was extracted for each sample, selected for mRNA by poly-A selection, and then fragmented to a mean length of ~120-180 base pairs. Strand-specific cDNA libraries were prepared and sequenced on an Illumina HiSeq 2000 at depths of 45-171 million paired end 100bp reads per sample.

Alignment and Transcriptome Analysis

Raw reads were mapped to the human genome sequence (hg19)[34] using Tophat v2.0.6[35] and PCR duplicates were removed. Some samples had a low percentage of unique

* Co-written with Sook Wah Yee and Aparna Chhibber

reads likely due to limited starting material. Transcript structure assembly was performed with Cufflinks (v.2.0.2)[6] on each sample for each tissue type. To control for differing sequencing depths between tissue types, and the variable number of samples analyzed for each tissue type, gene expression analysis was performed on a subset of the data: 20 million reads per sample and 18 samples per tissue type. Gene expression values (in Fragments per Kilobase of Exon Mapped, FPKM) were calculated by summing per-isoform FPKM values generated by Cuffdiff (v2.2.1)[6] for each sample or by tissue type. Throughout, gene estimates are used unless isoforms are specifically mentioned.

To discover novel splice events and analyze differential splicing, the subsampled reads were run through the JuncBASE v0.6[10] pipeline. JuncBASE uses junction reads from an RNA-Seq experiment to calculate inclusion and exclusion of individual splicing events. These are measured as percent spliced in (PSI). Such measures are generally more reliable than isoform reconstruction as they require less inference.

Validation

To validate selected splice events that were not found in the gene annotations, we created primers specific to the novel event and looked for amplification by PCR using pooled liver cDNA. To validate the PSI estimates derived from RNA-seq, PSI values for two common and previously annotated splice variants in *HMGCR13(-)*[13] and *LDLR4(-)*[13], were quantified by qPCR in LCLs (n=39) from the same RNA that was used to prepare the RNA-seq libraries. The PSI values for these two events in LCLs calculated by qPCR and RNA-seq were positively correlated with R^2 values of 0.43 and 0.5.

To validate the patterns of pharmacogene expression and splicing identified in this study, we analyzed data from the Genotype Tissue Expression (GTEx) Project (v4)[36]. Expression (RPKM, mapped reads per kilobase per million mapped reads) values per individual, per gene were downloaded from the GTEx portal (http://gtexportal.org) to study variability in gene expression and patterns of expression across tissues. Aligned reads were downloaded from SRA/dbGaP and run through the JuncBASE pipeline in the same way as was done for the PGRN data to compare differential splicing patterns between the two datasets and novel junctions identified in the PGRN dataset. Further details regarding all methodology are included in the supplemental methods.

Acknowledgements

This study is supported in part by the NIH Pharmacogenomics Research Network (PGRN) RNA Sequencing Project and grants U01GM61390, HL65962, U19HL069757, R01GM094418 and U01GM092666. AC was supported by U01GM61390 and GM007175. CEF was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. SEB received funding from Tata Consultancy Services.

References

1. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. *The New England journal of medicine* 2011; **364**(12)**:** 1144-1153.

2. Evans WE, McLeod HL. Pharmacogenomics--drug disposition, drug targets, and side effects. *The New England journal of medicine* 2003; **348**(6)**:** 538-549.

3. Mohamed S, Syed BA. Commercial prospects for genomic sequencing technologies. *Nature reviews Drug discovery* 2013; **12**(5)**:** 341-342.

4. Smith RP, Lam ET, Markova S, Yee SW, Ahituv N. Pharmacogene regulatory elements: from discovery to applications. *Genome medicine* 2012; **4**(5)**:** 45.

5. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* 2014; **30**(9)**:** 418-426.

6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR*, et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 2012; **7**(3)**:** 562-578.

7. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology* 2010; **11**(10)**:** R106.

8. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome research* 2012; **22**(10)**:** 2008-2017.

9. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* 2010; **7**(12)**:** 1009-1015.

10. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S*, et al*. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome research* 2011; **21**(2)**:** 193-202.

11. McCarthy JJ, McLeod HL, Ginsburg GS. Genomic medicine: a decade of successes, challenges, and opportunities. *Science translational medicine* 2013; **5**(189)**:** 189sr184.

12. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 2013; **45**(6)**:** 580-585.

13. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F*, et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* 2012; **22**(9)**:** 1760-1774.

14. Iyer L, Das S, Janisch L, Wen M, Ramirez J, Karrison T*, et al*. UGT1A1*28 polymorphism as a determinant of irinotecan disposition and toxicity. *The pharmacogenomics journal* 2002; **2**(1)**:** 43-47.

15. Tukey RH, Strassburg CP, Mackenzie PI. Pharmacogenomics of human UDP-glucuronosyltransferases and irinotecan toxicity. *Molecular pharmacology* 2002; **62**(3)**:** 446-450.

16. Wang D, Poi MJ, Sun X, Gaedigk A, Leeder JS, Sadee W. Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with

two regulatory variants modulate CYP2D6 activity. *Human molecular genetics* 2014; **23**(1)**:** 268-278.

17. Kim J, Zhao K, Jiang P, Lu ZX, Wang J, Murray JC*, et al*. Transcriptome landscape of the human placenta. *BMC genomics* 2012; **13:** 115.

18. Farkas MH, Grant GR, White JA, Sousa ME, Consugar MB, Pierce EA. Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC genomics* 2013; **14**(1)**:** 486.

19. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 2008; **5**(7)**:** 621-628.

20. Barton HA, Lai Y, Goosen TC, Jones HM, El-Kattan AF, Gosset JR*, et al*. Model-based approaches to predict drug-drug interactions associated with hepatic uptake transporters: preclinical, clinical and beyond. *Expert opinion on drug metabolism & toxicology* 2013; **9**(4)**:** 459-472.

21. Gandhi A, Moorthy B, Ghose R. Drug disposition in pathophysiological conditions. *Current drug metabolism* 2012; **13**(9)**:** 1327-1344.

22. Masereeuw R, Russel FG. Therapeutic implications of renal anionic drug transporters. *Pharmacology & therapeutics* 2010; **126**(2)**:** 200-216.

23. Lai Y, Varma M, Feng B, Stephens JC, Kimoto E, El-Kattan A*, et al*. Impact of drug transporter pharmacogenomics on pharmacokinetic and pharmacodynamic variability - considerations for drug development. *Expert opinion on drug metabolism & toxicology* 2012; **8**(6)**:** 723-743.

24. Huang RS, Duan S, Kistner EO, Zhang W, Bleibel WK, Cox NJ*, et al*. Identification of genetic variants and gene expression relationships associated with pharmacogenes in humans. *Pharmacogenetics and genomics* 2008; **18**(6)**:** 545-549.

25. Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, Chasman DI*, et al*. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature* 2013.

26. Wheeler HE, Dolan ME. Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. *Pharmacogenomics* 2012; **13**(1)**:** 55-70.

27. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF*, et al*. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics* 2012; **92**(4)**:** 414-417.

28. Montreal Heart Institute Pharmacogenomics Center (2013). PharmaADME.

29. U.S. Food and Drug Admnistration (2013). Table of Pharmacogenomic Biomarkers in Drug Labels.

30. Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N. Pharmaco-miR: linking microRNAs and drug effects. *Briefings in bioinformatics* 2013.

31. Ivanov M, Kals M, Kacevska M, Metspalu A, Ingelman-Sundberg M, Milani L. In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. *Nucleic acids research* 2013; **41**(6)**:** e72.

32. Gamazon ER, Skol AD, Perera MA. The limits of genome-wide methods for pharmacogenomic testing. *Pharmacogenetics and genomics* 2012; **22**(4)**:** 261-272.

33. Sissung TM, English BC, Venzon D, Figg WD, Deeken JF. Clinical pharmacology and pharmacogenetics in a genomics era: the DMET platform. *Pharmacogenomics* 2010; **11**(1)**:** 89-103.

34. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M*, et al*. The UCSC Genome Browser database: 2014 update. *Nucleic acids research* 2014; **42**(Database issue)**:** D764-770.

35. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**(9)**:** 1105-1111.

36. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**(6235)**:** 648-660.

37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM*, et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature gentics* 2000; **25**(1)**:** 25-29.

38. Morrissey KM, Stocker SL, Wittwer MB, Xu L, Giacomini KM. Renal transporters in drug development. *Annual review of pharmacology and toxicology* 2013; **53:** 503-529.

39. Rask-Andersen M, Masuram S, Schioth HB. The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual review of pharmacology and toxicology* 2014; **54:** 9-26.

40. Abernethy DR, Schwartz JB. Calcium-antagonist drugs. *The New England journal of medicine* 1999; **341**(19)**:** 1447-1457.

41. George AL, Jr. Recent genetic discoveries implicating ion channels in human cardiovascular diseases. *Current opinion in pharmacology* 2014; **15:** 47-52.

42. Oshiro C, Thorn CF, Roden DM, Klein TE, Altman RB. KCNH2 pharmacogenomics summary. *Pharmacogenetics and genomics* 2010; **20**(12)**:** 775-777.

43. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* 2008; **40**(12)**:** 1413-1415.

44. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M*, et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008; **321**(5891)**:** 956-960.

45. Li M, Jia C, Kazmierkiewicz KL, Bowman AS, Tian L, Liu Y*, et al*. Comprehensive analysis of gene expression in human retina and supporting tissues. *Human molecular genetics* 2014; **23**(15)**:** 4001-4014.

46. Webb A, Papp AC, Sanford JC, Huang K, Parvin JD, Sadee W. Expression of mRNA transcripts encoding membrane transporters detected with whole transcriptome sequencing of human brain and liver. *Pharmacogenetics and genomics* 2013; **23**(5)**:** 269-278.

47. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J*, et al*. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics : MCP* 2014; **13**(2)**:** 397-406.

48. Gremel G, Wanders A, Cedernaes J, Fagerberg L, Hallstrom B, Edlund K*, et al*. The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling. *Journal of gastroenterology* 2014.

49. Choi YS, Dusting GJ, Stubbs S, Arunothayaraj S, Han XL, Collas P*, et al*. Differentiation of human adipose-derived stem cells into beating cardiomyocytes. *Journal of cellular and molecular medicine* 2010; **14**(4)**:** 878-889.

50. Planat-Benard V, Menard C, Andre M, Puceat M, Perez A, Garcia-Verdugo JM*, et al*. Spontaneous cardiomyocyte differentiation from adipose tissue stroma cells. *Circulation research* 2004; **94**(2)**:** 223-229.

51. Makita N, Horie M, Nakamura T, Ai T, Sasaki K, Yokoi H*, et al*. Drug-induced long-QT syndrome associated with a subclinical SCN5A mutation. *Circulation* 2002; **106**(10)**:** 1269-1274.

52. Shuraih M, Ai T, Vatta M, Sohma Y, Merkle EM, Taylor E*, et al*. A common SCN5A variant alters the responsiveness of human sodium channels to class I antiarrhythmic agents. *Journal of cardiovascular electrophysiology* 2007; **18**(4)**:** 434-440.

53. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America* 2003; **100**(1)**:** 189-192.

54. Cropp CD, Komori T, Shima JE, Urban TJ, Yee SW, More SS*, et al*. Organic anion transporter 2 (SLC22A7) is a facilitative transporter of cGMP. *Molecular pharmacology* 2008; **73**(4)**:** 1151-1158.

55. Dahlin A, Geier E, Stocker SL, Cropp CD, Grigorenko E, Bloomer M*, et al*. Gene expression profiling of transporters in the solute carrier and ATP-binding cassette superfamilies in human eye substructures. *Molecular pharmaceutics* 2013; **10**(2)**:** 650-663.

56. Kobayashi Y, Sakai R, Ohshiro N, Ohbayashi M, Kohyama N, Yamamoto T. Possible involvement of organic anion transporter 2 on the interaction of theophylline with erythromycin in the human liver. *Drug metabolism and disposition: the biological fate of chemicals* 2005; **33**(5)**:** 619-622.

57. Kobayashi Y, Ohshiro N, Sakai R, Ohbayashi M, Kohyama N, Yamamoto T. Transport mechanism and substrate specificity of human organic anion transporter 2 (hOat2 [SLC22A7]). *The Journal of pharmacy and pharmacology* 2005; **57**(5)**:** 573-578.

58. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics* 2013; **138**(1)**:** 103-141.

59. U.S. Department of Health and Human Services FaDAF, (CDER) CfDEaR (2012). Guidance for Industry: Drug Interaction Studies — Study Design, Data Analysis, Implications for Dosing, and Labeling Recommendations.

60. Wempe MF, Lightner JW, Miller B, Iwen TJ, Rice PJ, Wakui S*, et al*. Potent human uric acid transporter 1 inhibitors: in vitro and in vivo metabolism and pharmacokinetic studies. *Drug design, development and therapy* 2012; **6:** 323-339.

61. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ*, et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006; **313**(5795)**:** 1929-1935.

62. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews Cancer* 2006; **6**(10)**:** 813-823.

63. Wernicke C, Hellmann J, Finckh U, Rommelspacher H. Chronic ethanol exposure changes dopamine D2 receptor splicing during retinoic acid-induced differentiation of human SH-SY5Y cells. *Pharmacological reports : PR* 2010; **62**(4)**:** 649-663.

64. Medina MW, Gao F, Naidoo D, Rudel LL, Temel RE, McDaniel AL*, et al*. Coordinately regulated alternative splicing of genes involved in cholesterol biosynthesis and uptake. *PloS one* 2011; **6**(4)**:** e19420.

65. Solier S, Barb J, Zeeberg BR, Varma S, Ryan MC, Kohn KW*, et al*. Genome-wide analysis of novel splice variants induced by topoisomerase I poisoning shows preferential occurrence in genes encoding splicing factors. *Cancer research* 2010; **70**(20)**:** 8055-8065.

66. Stormo C, Kringen MK, Lyle R, Olstad OK, Sachse D, Berg JP*, et al*. RNA-Sequencing Analysis of HepG2 Cells Treated with Atorvastatin. *PloS one* 2014; **9**(8)**:** e105836.

67. Vivarelli S, Lenzken SC, Ruepp MD, Ranzini F, Maffioletti A, Alvarez R*, et al*. Paraquat modulates alternative pre-mRNA splicing by modifying the intracellular distribution of SRPK2. *PloS one* 2013; **8**(4)**:** e61980.

68. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA*, et al*. Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America* 2003; **100**(4)**:** 1896-1901.

69. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY*, et al*. Mapping the genetic architecture of gene expression in human liver. *PLoS biology* 2008; **6**(5)**:** e107.

70. Daly AK. Using genome-wide association studies to identify genes important in serious adverse drug reactions. *Annual review of pharmacology and toxicology* 2012; **52:** 21-35.

71. Daly AK. Pharmacogenomics of adverse drug reactions. *Genome medicine* 2013; **5**(1)**:** 5.

72. Barrie ES, Smith RM, Sanford JC, Sadee W. mRNA transcript diversity creates new opportunities for pharmacological intervention. *Molecular pharmacology* 2012; **81**(5)**:** 620-630.

73. Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC*, et al*. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome research* 2011; **21**(8)**:** 1260-1272.

74. Warnes GR. gplots: Various R Programming Tools for Plotting Data. 2015.

75. Wickham H. *ggplot2: elegant graphics for data analysis*. Springer: New York, 2009.