Published in final edited form as:

Adv Exp Med Biol. 2016 ; 907: 1–28. doi:10.1007/978-3-319-29073-7_1.

Experimental and computational considerations in the study of RNA binding protein-RNA interactions

Eric L Van Nostrand¹, Stephanie C. Huelga¹, and Gene W Yeo^{1,2,3}

¹Department of Cellular and Molecular Medicine, Stem Cell Program and Institute for Genomic Medicine, University of California, La Jolla, CA, USA

²Molecular Engineering Laboratory, A*STAR, Singapore

³Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Keywords

CLIP-seq; HITS-CLIP; PAR-CLIP; iCLIP; Ribo-seq; BRIC_seq; RNA-seq; microarray

Background

The simplicity of the central dogma of molecular biology that information encoded by DNA is transmitted via RNA to proteins (the essential building blocks of cells) masks the complex regulatory networks involved at each step of this process. In eukaryotes, transcribed RNA molecules undergo a number of modifications that include pre-mRNA splicing, nucleotide editing and polyadenylation, all of which are tightly coupled to their sub-cellular localization, presentation and accessibility to ribosomal proteins for translation. At each of these steps the RNA may interact with a combination of one or more of the hundreds to thousands of RNA binding proteins (RBPs) present in humans, providing a means of regulating the fate of these RNAs in a tissue-, temporal-, or condition-specific manner ¹.

The importance of RNA binding proteins in controlling RNA processing has been exemplified by studies of individual RNA substrates for decades. Recent advances in microarray and high-throughput sequencing technologies have made it possible to not only identify individual RNA-protein interactions, but to identify RNA binding protein targets genome-wide in a single experiment. RNA targets can be directly bound and also regulated by the RBPs, bound but unaffected by the association of the RBP, or not bound but indirectly affected by the RBP (Figure 1). In this chapter, we will discuss recent efforts to both identify and predict *in vitro* and *in vivo* RNA binding sites, and to integrate these direct targets with transcriptome profiling experiments in order to obtain a full picture of RNA regulatory networks. In particular, we will discuss the computational challenges implicit in expanding from analysis of targets of a single RBP to integration of target information for dozens or hundreds of the estimated thousand RBPs in the human genome.

What is an RNA binding protein?

Indeed, the question of how many RNA binding proteins there are remains an open and challenging area of research. Most commonly studied RNA proteins interact with RNA through well-characterized RNA binding domains, including zinc finger (ZNF), RNA recognition motif (RRM), and helicase domains that can interact with RNA in various forms (e.g. unstructured, single-stranded, double-stranded or a combination of both). Using such protein domains to search the human genome yields estimates of ~600 RNA binding proteins¹. However, various examples of RNA binding proteins with non-canonical interactions have long suggested that other modes of interactions remain uncharacterized, suggesting that the number of true RNA binding proteins may be larger than expected.

Recently, efforts to identify novel RNA binding proteins using various approaches have further expanded the set of putative RNA regulators. In yeast, protein microarrays (in which a large number of proteins are spotted on a microarray and then probed with specific RNA transcripts of interest) identified a significant number of previously unannotated RNA binding proteins, including surprising interactions between a number of enzymes and RNA^{2,3}. More recently, an RNA pull-down followed by mass spectrometry approach in human cells identified nearly 900 proteins interacting with RNA in human HeLa cells and 555 in mouse embryonic stem cells, including more than 300 and 250 respectively that were not previously annotated as RNA binding proteins^{4,5}.

In addition to raising interesting computational questions in the area of predicting which proteins will interact with RNA and with what mechanism, these results raise the question of what truly defines an RNA binding protein. The single-stranded nature of RNA allows an RNA molecule to form a wide variety of secondary, tertiary, and even quaternary structures that enable a protein to not only interact with the ribonucleotide sequence, but also the nucleotide backbone in unique forms governed by the structure of the RNA. Indeed, RNA aptamers can be evolved *in vitro* that strongly interact with small molecules such as fluorophores⁶ or specific proteins of interest^{7,8}.

If an RNA can be engineered to specifically bind to GFP with high affinity, does this mean that GFP is an RNA-binding protein? Conversely, if an RNA- or protein-pulldown indicates an *in vivo* RNA-protein interaction, that would appear to be convincing evidence that the protein is an “RNA-binding” protein. However, the term “RNA-binding protein” has traditionally been reserved for describing proteins that functionally interact with RNA - i.e., that the protein’s interaction with RNA molecules causes some differential regulation of RNA processing in a regulated manner. Thus, as our abilities to detect *in vivo* protein-RNA interactions continue to improve, consideration will have to be given as to whether identification of RNA-protein interactions is sufficient, or whether it is more relevant to combine these results with analyses that also address whether this interaction with RNA drives the fate of the RNA.

Identification of RNA binding protein binding sites *in vivo*

A genome-wide view of RNA binding protein interactions is essential to understand how RNA binding proteins recognize RNA and drive differential RNA processing. These methods are generally referred to as “cross-linking followed by immunoprecipitation and high-throughput sequencing” or CLIP-seq, as a parallel to the ChIP-seq approaches used to purify DNA binding proteins and the associated bound chromatin. Many variations (HITS-CLIP, PAR-CLIP, iCLIP) have been described^{9–12}, all of which follow the same basic strategy: crosslinking of protein-RNA interactions, digestion or shearing of the RNA to smaller fragments, immunoprecipitation of the targeted protein (and associated RNA) using a target-specific antibody, isolation of RNA, and ligation of linkers and RT-PCR amplification to generate libraries compatible with either microarrays or, more recently, high-throughput sequencing.

Various algorithms have been developed to identify regions of significant association from CLIP-seq experiments. Fundamentally, these algorithms attempt to solve the question of whether the read density within a given region is characteristic of a binding site or characteristic of random association. Early methods simply counted the number of reads within a defined window size (e.g., 100 nucleotides) and asked whether the number of reads was significantly enriched above expected, given the number of total reads throughout that transcript. More recent methods have incorporated additional features, such as improved statistical modeling of read distributions, information about repetitive elements and read mappability, and the canonical ‘shape’ of a binding site (the local distribution of reads around a binding site) to improve identification of true binding sites^{9,13}.

Depending on the experimental methodology used, other information can also be incorporated into binding site identification. Using general CLIP-seq techniques, typical cluster sizes are ~50–150 wide, depending on the size selection used during library preparation. Recognizing that the reverse transcription step was often inefficient and terminated at the nucleotide cross-linked to the RNA binding protein, König *et al* (2010) exploited this truncation to anneal a linker oligonucleotide at this terminal end. Termed ‘iCLIP’, this approach led to a pile-up of reads at the cross-link site, enabling single-nucleotide resolution identification of binding sites⁹. Using more traditional CLIP datasets, Zhang & Darnell recognized that single-nucleotide resolution could also be achieved computationally by searching sequencing reads for single nucleotide insertions and deletions characteristic of reverse transcriptase skipping at the cross-link site¹⁴. Although less than 20% of reads showed such mutations, this density was sufficient both to identify the bound residues at most binding sites, as well as to generate insights into binding motifs and local sequence structure at binding sites genome-wide¹⁴.

Before cluster identification, an important step is quality control to determine whether the CLIP-seq experiment was successful and generated reliable data. As immunoprecipitation of RNA binding proteins (and subsequent adapter ligation onto RNA) is often inefficient, CLIP-seq libraries are often amplified from small amounts of RNA input, leading to significant concerns about PCR amplification biases when a high number of PCR cycles are needed to amplify enough DNA for sequencing. Indeed, a high degree of read redundancy

(multiple reads of identical sequence) is typically observed in CLIP-seq experiments, up to 90% in many published datasets (*Van Nostrand EL, Pratt G, Yeo GW, unpublished*). However, as the identified clusters are typically short (<200 nt) and (as described above) often start at the cross-linked nucleotide, it is impossible to distinguish PCR amplification from multiple unique fragments using the standard CLIP methodology, and most approaches will simply compress these multiple reads into one for downstream analysis. Paired-end sequencing can address some of these concerns (as the odds are lower that two unique fragments would have the same start and end position). iCLIP provides further ability to distinguish these possibilities by incorporating a random five base sequence (along with a multiplexing barcode) into the ligated RNA adapter that will ultimately be sequenced in each read⁹. After sequencing, computational processing can be performed to identify the random adapter sequence, barcode, and RNA fragment sequence. Using this approach 1024 (4⁵) reads can map to the same location instead of only one, significantly improving signal to noise at true binding sites.

Challenges of peak finding for RNA binding proteins compared to DNA binding proteins

It is appealing to turn to the field of DNA binding proteins, as the identification of *in vivo* binding sites for DNA-binding transcription factors requires solving a number of similar computational challenges. Akin to identifying RNA binding protein binding sites, the experimental association of transcription factors to their DNA binding sites involves cross-linking the factor to its genomic DNA (typically using formaldehyde), fragmenting the DNA, immunoprecipitating the factor and isolating the associated 100–150 bp DNA fragments, and then quantifying the associated DNA using high-throughput sequencing (chromatin immunoprecipitation followed by sequencing, or ChIP-seq)¹⁵.

Although the fact that it is RNA instead of DNA being queried seems like a trivial difference, in reality there are substantial challenges distinct from identifying DNA binding sites. In a ChIP-seq experiment in normal tissues, the expected read density across the genome is roughly equal (as all DNA regions are expected to be present on each of the two chromatids in each cell); although this genome-wide equivalence may not hold for all cell lines (which may have acquired chromosomal aberrations), the distribution in a reasonably large local region around the binding site should show this equivalency. In contrast, as each RNA transcript is expressed at a different level, all calculations to identify RNA binding sites are limited to within-transcript read information. RNA splicing also represents a complication, as introns (which are among the most common binding sites for many RNA binding proteins, and represent the functional binding sites for many splicing regulators) are rapidly spliced from most transcripts, meaning that only a small fraction of messages for a gene of interest will still contain a potential RNA binding protein - intronic sequence interaction that can be quantified. Additionally, the presence of RNA requires additional, more complicated ligation steps to generate tagged RNA molecules that can be reverse transcribed into cDNA. Finally, because a transcription factor associates with double-stranded DNA, DNA binding sites obtained by ChIP-seq will be identified by sequencing reads that map to both the Watson and Crick strands. However, the distributions of these

reads on the two strands will be separated by a width equal to the size to which the DNA is sheared, enabling this characteristic shift to be used in downstream algorithms as both a validation of true binding sites as well as a validation of proper library construction ¹⁶.

However, there are also more subtle differences. Although ChIP-seq experiments often can use a non-immunoprecipitated 'input' sample or an 'IgG-only' secondary antibody pull-down with no primary antibody as a negative control, those negative control samples have proven more difficult to obtain for CLIP-seq experiments. Due to the methodology of CLIP-seq, an input sample would introduce various biases that would overwhelm the ability to use it as a negative comparison, and attempts to leave out the primary antibody have typically yielded too little RNA to yield useful high complexity libraries for sequencing. In addition, the vast array of RNA messages in the cell can include overlapping but distinct genes, genes with multiple isoforms, and unannotated transcripts, which all add complexity to analysis by making it difficult to unambiguously assign reads to individual transcripts. Because of this, most CLIP-seq analysis is still done using custom-written pipelines by individual labs, and has not yet coalesced into one generally used software package across the field.

RBP-responsive RNA targets

Identification of the *in vivo* binding sites for an RNA binding protein can give insight into the molecular mechanism of binding for that factor. However, to understand the molecular basis for the phenotypes observed upon knockdown or overexpression it is necessary to identify not only those targets directly bound by the RBP, but also those that show RBP-dependent altered RNA processing. Individual target and deep mutational analysis of minigene-based assays have provided significant insights in the study of individual RNA binding protein responsive targets over the past decades ¹⁷. More recently, it has become possible to modulate an RNA binding protein and query the effect on various outputs (RNA expression, alternative isoform usage, and ribosomal occupancy) in a genome-wide manner. However, inherent in each of these assays are various experimental decisions and computational challenges that should be considered during the initial planning stages.

Choosing between depletion versus over-expression experiments

The first decision when designing the experiment is whether to profile loss-of-function or gain-of-function of the protein of interest. It is ideal to mutate both the maternal and paternal alleles of the protein to generate a complete loss of function, as this ensures that the function of the protein is completely lost. However, creating such a null mutation can often be difficult and require significant investment of resources, particularly in whole organisms. More common is to knock down the protein, which can be performed using a variety of techniques that depend on the cell-type and duration of knockdown desired ¹⁸. The commercial availability of knockdown reagents such as siRNA or shRNAs for nearly all protein-coding human and mouse genes, as well as the ease of generating and using these reagents, makes knockdown both the simplest as well as the most scalable experiment. However, there exist a number of caveats with knockdown experiments that are important to consider in downstream analyses. First, most knockdown reagents (including shRNA and siRNA molecules) use an antisense oligonucleotide which leads to RNA degradation

through the RNA interference pathway, but which can have uncharacterized off-target effects if the oligonucleotide binds to other expressed genes. As such, it is important to perform multiple independent knockdowns with antisense oligonucleotides targeting multiple regions of the gene, in order to separate true RBP targets from off-target effects¹⁹. Secondly, many RNA binding proteins are found in families with multiple highly similar paralogs that may have functional redundancy. As such, knockdown of only one family member may yield few if any altered targets, which will only be revealed after simultaneous knockdown of multiple family members. Various degrees of redundancy have been observed for a variety of RNA binding proteins, including the Muscleblind family²⁰ and RBFOX family^{21,22}, indicating that this may represent a concern for analysis of many RBPs.

Ectopic expression of an RNA binding protein can avoid this issue of functional redundancy, as well as avoiding concerns for off-target effects. Additionally, it is particularly powerful when the RBP of interest is introduced into a cell line in which it is not normally expressed, as this may help to isolate the signal of true targets of the RBP in isolation of other RBPs that it often interacts with. However, the expression of a transgene brings with it its own set of concerns. Significant over-expression of DNA- and RNA- binding proteins can lead to binding to (and subsequent regulation of) lower-affinity binding sites that would not be physiologically relevant *in vivo* targets²³. Thus, it becomes important to design the promoter, 3' UTR, and other aspects of the transgene in such a way to express the transgene at roughly equivalent levels to its expression in other tissues. This can also be addressed computationally, by comparing RBP-responsive targets against other information to verify whether the targets observed are likely to be relevant under normal conditions. Depending on the experimental method, over-expression may also require selection of an individual isoform of the RNA binding protein if the entire gene region is too large to reliably transmit into the cell line of interest. As multiple isoforms of an RBP can show distinct regulatory activity and even differential sub-cellular localization²⁴, it becomes critical to consider whether the activity of the gene or only of that specific isoform is being read out.

Quantitation of RNA isoform abundance

The second major decision in identifying RNA binding protein targets is what aspect of RNA processing will be queried, and how altered targets will be identified. Currently, much effort has been devoted to identification of alternative splicing events, which can be quantified using both high-throughput sequencing (RNA-seq) and microarray approaches. In both cases, the goal of these analyses is to quantify the percent of isoforms that contain an alternative exon (quantitated as the percent spliced in, or Psi (Ψ) value). However, the methodological and technical differences between these two technologies mean that the computational challenges for analysis, and the algorithms and approaches used to solve these challenges, are somewhat unique for each platform. Microarrays are best suited for quantitation of previously identified alternative splicing events, whereas RNA-seq is able to identify alternative isoforms *de novo*. For both, there are two basic approaches to determine psi values: a direct approach, in which one typically looks for signal from the specific exon-exon junctions that correspond to various isoforms, or an indirect approach, where the level of each isoform is computationally inferred based on not only this direct information but also by modeling to predict inclusion in situations that lack direct evidence. Given infinite

information, direct approaches are generally preferred; however, there are many cases where indirect approaches can either help gain insights with lower sequencing depth or can be used in cases where direct information is not possible.

Identification of altered RNA splicing events by microarray

Microarrays contain thousands of short oligonucleotides that are each complementary to a specific target sequence, ranging from 25-mers on Affymetrix platforms to 60-mers using Agilent arrays. These oligonucleotides are then spotted to a surface with a known pattern, and a fluorescently labeled cDNA library is flowed over the array surface. DNA fragments complementary to the spotted oligonucleotides will anneal at that position, and the fluorescence of each 'spot' can be used as a read-out for expression level of the sequence complementary to the oligonucleotide probe spotted at that position ²⁵.

With development of high density microarrays, it is now common for the microarray to contain multiple probes that tile across a region of interest, providing multiple independent readouts of expression of an individual gene, exon, or splice junction (Figure 3a). For quantitation of RNA processing, splice junction probes are particularly helpful as they can be used to directly measure expression of individual alternative isoforms. Thus, a cassette exon being queried will typically have 6 'probe sets' of interest (each of which contain multiple individual probes): one probe set tiling the exon (represented as ASexon in Figure 3a) as well as one tiling each of the upstream and downstream exons (UPexon and DNexon), and three splice junction probe sets tiling each of the three splice junctions (UPjunction, DNjunction, and ASjunction) (Figure 3a). Quantitation of exon inclusion (Ψ) can then be performed by comparing the signal at probes supporting exon inclusion (probe sets ASexon, UPjunction, and DNjunction) against those supporting exon exclusion (probeset ASjunction) ²⁶. Quantification of other types of alternative splicing simply requires different probe positioning to measure the novel junctions generated by those events. Figure 3c shows representative data for SLK exon 13 in experiments profiling differentiation of induced pluripotent stem cells to motor neurons. Notably, while upstream and downstream constitutive exons showed stable expression between iPSC and differentiated neuron samples, significantly decreased signal is observed at all three probesets representing exon inclusion. Concordantly, signal was increased at the ASjunction probe that directly measures exon exclusion.

Analysis of these microarrays can be challenging, with multiple methods described that incorporate these multiple probe sets into one value that characterizes differential processing between samples ^{26,27}. In the simplest approach, intensity at probesets targeting neighboring constitutive exons can be used to normalize for sample or gene expression differences, and performing a t-test between the three knockdown and three control replicate samples can reveal significantly altered individual exons ²⁸. The addition of probesets covering the splice junctions makes this calculation more robust by providing additional independent measurements that should all show the same fold-change between conditions ^{26,27,29,30}. This has led to the development of probabilistic models that can infer the underlying expression of individual isoforms from probe intensity measurements (reviewed in *Chen, L. 2011*) ³¹.

Quantifying alternative splicing by high-throughput sequencing

Identification of alternative splicing by high-throughput sequencing (RNA-seq) is performed by starting with RNA from a sample of interest. As total RNA samples contain mostly ribosomal RNA, a purification to deplete rRNA is typically performed by either selection with poly-T oligonucleotides to isolate only polyadenylated RNA (enriching for mRNA), or by selective depletion of ribosomal RNA using bead-coupled antisense oligonucleotides (using kits such as the Ribo-Zero kit available from Illumina). RNA is then fragmented, reverse transcribed into cDNA using random hexamer primers, and the second cDNA strand is synthesized. Various methods can be used to preserve strand information; one commonly used technique involves replacing dTTP with dUTP in the second strand synthesis step, and then performing PCR with a polymerase enzyme that is blocked at dUTP nucleotides³². Oligonucleotides containing adapters for high-throughput sequencing are then ligated, and the sample is PCR amplified and size-selected for sequencing. Sequencing a sample using one lane of Illumina HiSeq 2500 or 4000 machines (the current methodology with greatest sequencing depth) yields over two hundred million reads, each of which can be 30–250 bases and in either single-end or paired-end format (depending on user selection of sequencing kits). Reads are then mapped to the genome (typically using one of several publicly available programs^{33,34}), and post-processed along with a genome annotation to derive position- as well as transcript-level read density profiles³⁵.

Figure 2 shows the type of data utilized for two types of alternative isoforms: cassette exons, and alternative polyadenylation sites. For cassette exons, reads that map to the junction between the upstream and downstream exon provide strong, direct evidence of exon exclusion. In contrast, reads that map to the exon itself or that map to exon-exon junctions with the upstream or downstream exon provide evidence of exon inclusion. The ratio of inclusion reads (normalized for the increased mappable sequence length) relative to exclusion reads will then provide a ratio of exon inclusion in a sample of interest. A simple hypergeometric or Pearson's Chi-square test between exclusion and inclusion reads can then be used to determine whether an event shows altered splicing across two samples³⁶. The simplicity of this approach, combined with the fact that all reads used in the analysis provide direct evidence either for an inclusion or exclusion isoform, make it ideal for experiments with extremely high sequencing depth (or for highly expressed transcripts). However, as exclusion can only be identified by reads that span the specific exclusion exon-exon junction, short read length will often lead to few reads that actually span this junction, limiting detection.

This direct quantitation of Psi values does not use any information about reads that map elsewhere in the transcript. However, expression of flanking exons can provide useful information - for example, if the upstream and downstream exons each show dramatically higher read densities than a central queried exon, this would suggest that the exon is likely excluded in some percent of transcripts. Although this drop in read density is only probabilistic evidence of an excluded exon and not direct observation of the excluded isoform, inclusion of such information provides more accurate estimates of exon inclusion rates³⁷. Because isoform quantitation using this approach is now no longer deterministic (based on read counting) but now requires estimation of the most likely Psi value that would

lead to the observed read densities, algorithms that incorporate such information estimate these values using various machine learning and probabilistic approaches. MISO, one of the more popularly used software packages for this analysis, treats this estimation as a Bayesian inference and uses the observed read densities across the queried exons and exon-exon junctions as a ‘posterior’ that can be used to infer the ‘prior’, or true (but unobserved) Psi value³⁷. The estimated Psi values, along with confidence intervals, can then be used to identify significantly altered alternative splicing between two tissues or conditions. Further development of other methods to robustly estimate Psi values should yield further improvements in quantitation of splicing changes by RNA-seq.

Because isoform quantitation by RNA-seq depends on sequencing reads that specifically map to the queried exon (and flanking exons and splice junctions), the total number of reads obtained from high-throughput sequencing represents a major consideration for analysis success. Detection of an alternative event thus depends on two factors in addition to the splicing event itself: overall sequencing depth, and expression of the entire gene. Splicing events in highly expressed genes will be easy to detect even with low-coverage experiments (for example, alternative intron retention in the 3’ UTR of NAP1L1 can be easily detected in less than 10 million reads, achievable in a single run on an Illumina MiSeq machine). In contrast, alternative events in lowly expressed genes may be difficult to detect using more complex algorithms and the hundreds of millions of reads obtained in more deep sequencing on an Illumina HiSeq machine. Thus, analysis of RNA processing by RNA-seq will always involve a trade-off between sequencing depth (and thus, cost) and the ability to detect events of genes with lower expression (which may include biologically relevant events in DNA- or RNA-binding proteins that are not always highly abundant even when functional).

Identification of novel alternative splicing events by RNA-seq and microarray

Figures 3 and 4 describe how the expression of a previously identified alternatively spliced exon can be quantified. However, it is also possible to adapt these techniques for *de novo* identification of alternative isoforms, finding previously unseen isoforms that may be unique to specific cell-types or conditions. As described above, sequencing reads that perfectly map across a splice junction can be used to verify a novel alternative event; thus, identification of novel alternative events requires using sequencing reads that do not map to the annotated genome or transcriptome to infer the splicing event that must have occurred to generate the observed sequence. Although this is easiest with paired end reads, longer read length in high-throughput sequencing has improved the ability to detect novel splicing events using single-end reads as well.

Because of this additional step, algorithms that not only align reads but also perform *de novo* transcript assembly and isoform identification have typically been slower and required heavier computational resources than simple mapping approaches. The TopHat and Cufflinks suite of programs have been widely used to identify and quantify novel splicing events, but can require significant resources for large-scale datasets³⁵. To handle the scale of RNA-seq data generated by the ENCODE consortium, the alignment program STAR was

developed³³. In STAR, the largest 5' fragment of the read that maps to the genome is identified first. Then, the mapping is repeated for any remaining unaligned sequence, identifying split reads that identify splice junctions or other fusion products. Further processing can then identify the exons (whether annotated or novel) that these split reads are derived from. Although STAR requires considerable computational resources (~40gb of RAM for mapping to the human genome, although this can be relaxed to ~16gb at a trade-off to mapping speed), it enables rapid identification of novel alternative splicing events from large compendia of RNA-seq datasets³³. Other algorithms use similar approaches to identify spliced reads, with various tradeoffs for speed and sensitivity.

Identification of putative alternative exons from microarrays is more challenging, but can be done in different ways depending on the microarray strategy employed. As described above, direct identification of an alternative isoform by microarray requires probes designed to span the exon-exon junction created by that alternative event. Thus, direct identification of novel alternative events requires tiling all exon-exon junctions (for example, probes that span the junction between exons 1 and 2, 1 and 3, 1 and 4, 1 and 5, etc.) in order to identify signals from previously unannotated splice events. Profiling of 52 tissues using such an approach identified thousands of novel alternative splicing events with a high accuracy rate²⁵; however, such custom array designs are not commonly used due to cost, complexity of analysis, and the difficulty in balancing exhaustive coverage to identify novel events with wasted coverage of non-observed junctions. Indirect inference of novel splicing events can also be performed using more common exon-level microarrays, including the Affymetrix Human Exon arrays. As was the case with RNA-seq data, indirect identification of novel exon exclusion events can be inferred by comparing relative expression of an exon, as well as its upstream and downstream flanking exons, across multiple samples. However, this approach has lower detection ability and is limited to events that are assayed on the array, typically limiting analysis to simple exon skipping events.

Thus, RNA-seq and microarray based approaches each have independent advantages and disadvantages to identification and analysis of alternative RNA processing events. Microarrays provide substantial advantages to quantification of a known set of events, as all events can be easily quantified genome-wide in a single experiment. In contrast, the level of expression of each gene represents a significant hurdle in quantifying RNA processing by RNA-seq, as ever-increasing amounts of sequencing are required to detect alternative isoforms in lowly expressed genes. Recent methods of generating targeted RNA-seq libraries to enrich for specific desired sequences can alleviate this concern somewhat, but add significant cost to the experiment. In contrast, RNA-seq is uniquely well-suited to identify novel alternative events, as it simply requires the development and implementation of computational tools to identify these events from a standard RNA-seq library. In contrast, identification of novel events by microarray requires either significant alteration to the design of the microarray itself, or specialized analysis tools that can detect exon skipping but are limited in their ability to detect more complex alternative splicing events. Thus, the methodology being chosen will depend heavily on the experimental design - RNA-seq is ideal when profiling a novel tissue or cell-type, or manipulation of an RNA processing factor that may have unknown roles, in order to identify novel and complex alternative events that may have been previously missed. In contrast, large-scale profiling of known splicing events

is perhaps better suited for microarray-based approaches, as they will provide the most robust coverage of all events profiled on the microarray.

Alternative polyadenylation sites

Recent publications have identified alternative polyadenylation as a significant factor in functional isoform diversity, as the generation of isoforms that contain or lack an extended 3' untranslated region can lead to alternative inclusion of microRNA target sites or other regulatory sequences that drive differential RNA regulation. Multiple groups have developed techniques to identify polyadenylation sites by RNA-seq, by identifying sequencing reads which contain a region which maps to the genome, and either terminate in a string of A nucleotides or begin with a string of T nucleotides that do not map (and are thus characteristic of the poly-A tail), or by selecting for only these 3' end fragments during either library preparation or sequencing^{38,39}. Differential alternative poly-A site usage can then be identified using standard t-test or Fisher's Exact tests. These direct approaches provide the strongest evidence for alternative polyadenylation; however, it can also be inferred by looking for a characteristic drop in read density at a specific point in the 3'UTR. Similar to analysis of alternative splicing events, programs such as MISO can use read density before and after a putative alternative polyadenylation site to estimate the percent usage of the two sites³⁷.

RNA stability

Once an RNA transcript has been spliced, polyadenylated, and exported from the nucleus, the half-life of each mRNA molecule is tightly controlled through regulatory interactions that involve various RNA binding proteins. These interactions include targeted degradation mechanisms, such as microRNA-mediated RNA degradation through interactions with the Argonaute RNA binding protein family and nonsense-mediated decay of transcripts containing premature stop codons that involves the activity of RBPs including UPF1, as well as more subtle regulation of RNA stability^{40,41}. AU-rich elements, one general class of RNA sequence element commonly found in 3'UTRs, can be bound by various factors that either stabilize the mRNA (including the HuD/HuR family of RBPs) or destabilize the RNA (including factors like AUF1)^{41,42}.

Given an RNA binding protein that regulates stability of a particular mRNA, the effect of knocking down that RBP will be to either increase (for destabilizing interactions) or decrease (for stabilizing interactions) the observed expression level of that mRNA. Thus, a simple gene expression quantification (either individually by qPCR, or done in high-throughput by RNA-seq or microarray) can serve as a first-pass analysis to identify the targets of an RBP regulated at the level of message stability. However, this analysis will also identify numerous differentially expressed genes that are secondarily regulated by altered transcription factors or other regulators and not true targets of the RBP itself. To address this limitation, specialized protocols like the BRIC-seq methodology have been developed to specifically quantitate RNA half-life⁴³. By metabolically labeling RNA with 5'-bromo-uridine (BrU) followed by a chase with fresh media, only RNA molecules transcribed until a specific timepoint are labeled. RNA is then sequenced at specified timepoints (0hr, 4hr, 8hr,

and 12hr) post-chase to quantify the decrease in BrU incorporation level over time, which serves as a proxy for RNA half-life. Application of this method in HeLa cells identified significant differences in gene function among genes with long and short half-lives, suggesting that RNA stability may play critical roles in regulating specific cellular functions⁴³.

Ribosome occupancy

Recent technical advances have also enabled efforts to quantify altered regulation of translation. The ideal experiment, full proteome mass-spectrometry to quantify expression of all proteins, remains technically challenging and cost-prohibitive for most groups. However, methods to profile actively translating ribosomes have shown that ribosomal occupancy can serve as an effective proxy for translation rate, if not for protein expression itself^{44,45}. Briefly, RNA is digested and 80S ribosomes (along with associated RNA fragments) are isolated by centrifugation through a sucrose cushion. Protected RNA fragments are then purified, ligated to oligonucleotide linkers, and amplified to generate libraries compatible with high-throughput sequencing⁴⁴. This methodology is now available through the commercial ARTseq Ribosome Profiling Kit from Epicentre, making it possible for labs without previous expertise to characterize ribosomal occupancy genome-wide in a standardized manner. Further specialized protocols have made use of the biotin ligase protein BirA to specifically tag ribosomes localized proximal to specific subcellular structures, allowing profiling of localized translation [Jan weissman].

However, some specialized computational analysis is required to take full advantage of ribosome profiling data generated by high-throughput sequencing. At the core, analysis of these datasets faces similar challenges to CLIP-seq datasets; although profiling of translation is both experimentally as well as computationally more challenging than RNA profiling (by RNA-seq or microarray), it has two major advantages. First, numerous studies have suggested a surprisingly low correlation between mRNA and protein expression (whether raw levels or differential expression between two conditions)^{46,47}, suggesting that profiling that is closer to protein expression itself will yield results that will correlate more closely with true biology. Second, recent studies have revealed that individual ribosomal subunits can show tissue-specific expression patterns and can play very specific roles in translation regulation. For example, large ribosomal subunit RPL38 appears to specifically regulates translation of HOX mRNAs, as mutation of RPL38 in mice has little effect on synthesis of other proteins but leads to significant developmental phenotypes characteristic of loss of HOX protein expression⁴⁸. Thus, analysis that focuses solely on mRNA expression levels may miss significant translational regulation with significant roles in development and disease.

Challenges of scale

Until recently, the approaches described above were largely used to analyze either single or small numbers of datasets. However, the combination of ENCODE-scale projects as well as the continual decrease in sequencing cost has enabled projects that profile RNA processing in hundreds, or even thousands, of conditions. Indeed, preparation of RNA-seq libraries in

96-well multiplexed format is now standardized in kit format from various commercial vendors. This increase in scale represents significant challenges; analysis pipelines that require on the order of one day per dataset do not represent a problem for small-scale experiments, but would be prohibitively long for sequential analysis of thousands of datasets. Thus, one of the major current computational challenges facing RNA analysis is to reformulate current approaches in ways that are faster and more rapidly scalable in order to handle large numbers of datasets.

Progress on this front is being made for many aspects of the various analysis pipelines. The availability of large computational clusters, and even commercial availability of multiple core desktop computers with 32–64gb of RAM, have placed CPU time and storage space as the rate-limiting steps for primary dataset analyses. The STAR alignment program (developed as part of the ENCODE RNA-seq efforts) has made substantial progress in addressing read mapping, typically one of the slowest and most resource-intensive steps in RNA-seq analysis, increasing mapping speed by a factor of 50 above pre-existing methods at a cost of increased RAM usage that is still reasonably available in standard computer facilities³³. Samtools, implementing the SAM and companion binary compressed BAM file formats, has enabled rapid post-mapping read statistics and alignment retrieval with a highly compressed file format that removes the need to store multiple large intermediate read alignment files⁴⁹. Further work remains to optimize downstream analyses, such as CLIP-seq cluster identification, to increase the scalability of these resources as well.

The second major aspect of scaling is the automation of standard quality control and other basic analyses. With thousands of datasets, it is impractical to deeply explore each dataset to look for indications of contamination, low-complexity or over-amplified sequence libraries, or other potential biases. Thus, these large-scale analyses must include some level of automatic processing to flag potentially troublesome datasets for further analysis. In addition to basic technical quality control performed during high-throughput sequencing, programs such as FastQC incorporate a variety of calculations including per base and per read quality score, sequence bias, and over-represented k-mer analyses to flag potentially problematic datasets⁵⁰. Efforts to optimize analysis-specific measures, such as the percent of CLIP-seq reads that fall into identified binding site clusters or the distribution of mismatched bases in iCLIP (that indicate the position of binding at single-base resolution) are ongoing.

Finally, efforts continue to develop methods of visualizing both quality control metrics as well as processed data from multiple datasets in parallel. FastQC provides useful tools to visualize the various quality control metrics, but developing ways to combine these reports into easy to view cross-sample reports remains an open challenge. RNA-SeQC, developed by the Cancer Genome Atlas, takes in highly compressed BAM files and provides a more comprehensive set of quality control metrics as both HTML and tab delimited files, helping rapid analysis of large numbers of datasets in parallel⁵¹. Similarly, visualizing large-scale post-analysis data in a useful manner remains challenging. Tools like the locally installed Integrated Genomics Viewer (IGV)⁵² and web-based resources like the UCSC Genome Browser⁵³ provide simple frameworks in which multiple datasets can be viewed and compared. However, simultaneous viewing of many datasets (each of which could be gigabytes in size) is somewhat more challenging, depending on the ultimate goal. With

sufficient computational resources, local installation of IGV (or similar browsers) is ideal for individual viewing, but lacks easy sharing across groups. In addition to the standard UCSC Genome Browser hosted by UCSC (which has limits on the number and size of datasets that can be uploaded and directly viewed), UCSC has also provided a mechanism to install the browser on a local server, providing a way to avoid data transfer issues by accessing private locally hosted datasets. Additionally, UCSC and others have recently developed a “Track Hub” system, by which groups can visualize a large number of genome-scale datasets in the normal UCSC browser framework while the raw data is hosted on a remote server. In this way, large-scale datasets can be viewed alongside other public data (including the ENCODE data that is publicly hosted at UCSC) while still maintaining data privacy and local storage.

Learning predictive RNA processing networks

The identification of RBP binding sites (both *in vitro* and *in vivo*) can provide molecular insights into the basic mechanisms of RBP binding and function. However, it is infeasible to identify the targets of every RBP independently in all the various cell-types, tissues, and conditions in which the RBP may function. Thus, computational techniques need to be developed that can take target information generated in easy-to-manipulate cell lines and tissues, and integrate this information into predictive models of RBP regulation which can then be used to predict roles of an RBP in conditions not yet experimentally explored. Although challenging, research into these models has progressed along three tracks: first, generating predictive models of splicing by using sequence information alone; second, by deeply exploring the targets of an individual RBP, and third, by combining target information for multiple RBPs to infer combinatorial effects of multiple RBP regulators.

The ability to predict whether mutations or polymorphisms are likely to alter splicing, and ultimately isoform usage ratios themselves, would greatly aid rapid interpretation of potential disease-causal mutations identified in genome-wide association and whole-genome sequencing studies. Before it was experimentally possible to identify *in vivo* binding sites in a high-throughput and genome-wide manner, sequence motif enrichment and cross-species conservation provided the best opportunity to infer differential RNA regulation. By analyzing exonic and intronic sequences for motifs that are enriched nearby alternatively spliced exons, that showed unusual conservation across species, or that were enriched near splice sites, numerous efforts have identified a variety of short sequence motifs that correlate with splicing regulatory activity^{28,54}. Initially, these motifs (as well as other sequence features) were used to address a simple question: could alternatively spliced exons be distinguished from constitutively spliced exons based solely on sequence information? Using various machine learning models, various groups were able to perform such prediction with high accuracy in both mammals (human & mouse) as well as *Drosophila*, indicating that substantial regulatory information was contained within the RNA sequence itself^{55–57}.

Barash, *et al.* (2010) extended this analysis to ask whether such sequence features could be used to not only predict whether an exon would be alternatively spliced or not, but also in which specific tissue such alternatively splicing would occur. Using a machine learning approach with 1014 different sequence features (including binding motifs for known factors

as well as computationally identified motifs with yet unknown regulators, transcript structure features such as exon and intron size, etc.), Barash, *et al.* showed impressive accuracy in predicting tissue-specific alternative splicing patterns, including an ability to predict altered tissue-specific isoform usage upon mutation of putative regulatory elements⁵⁸. Xiong *et al.* (2015) extended this approach to the more basic question of predicting Psi values themselves by incorporating 1393 sequence features into a machine learning splicing predictor⁵⁹. This method showed good accuracy for prediction of Psi values (R^2 of 0.65, with higher accuracy for predictions of increasing confidence), sufficient to perform the first global analysis of the effect on splicing for all annotated single-nucleotide variants, identifying over twenty thousand common and rare variants predicted to significantly alter splicing regulation. The results of this and other efforts suggest that an ability to rapidly and accurately predict the effect of novel mutations on splicing will be achieved in the near future^{58–60}.

Integrating target information to generate regulatory maps for individual RBPs

Second, detailed analyses of the targets of individual RNA binding proteins can give an in-depth view of the differential regulatory activities an RBP can have in different contexts. Perhaps the most well-studied example of detailed characterization of the differential effect of RBP interaction at various positions in exons and flanking introns is the activity of neuronal splicing factors NOVA1 and NOVA2. NOVA proteins show specific expression patterns in the brain, and are essential for postnatal motor neuron survival^{61,62}. Various early molecular studies, including *in vitro* binding assays and X-ray crystallography, indicated that the KH-type RNA binding domains of NOVA directly interact with YCAY sequence motif clusters^{62–64}. Although alternative splicing of many exons were known to be regulated by NOVA, NOVA induced inclusion of some but exclusion of others, suggesting that a simple presence of YCAY clusters was not sufficient to predict the mode of NOVA regulation.

To better understand NOVA-mediated splicing regulation, Ule. *et al* (2006) developed the concept of a ‘splicing map’, in which they calculated the frequency of YCAY clusters at various positions within NOVA-dependent exons as well as their flanking introns⁶⁵. The degree of enrichment was then plotted on a generic, size-normalized exon-intron-exon structure to create two separate motif-enrichment plots: one for NOVA-dependent included exons, and another for NOVA-dependent excluded exons. These maps showed striking patterns of differential regulation by NOVA based on YCAY location; YCAY clusters within the alternative exon or located near the 5’ splice site of the upstream intron correlated with silencing activity, whereas YCAY clusters near the 5’ or 3’ splice sites in the downstream exon correlated with enhancing activity. Although this analysis was performed using sequence information alone, later work using *in vivo* binding sites for NOVA identified by a CLIP-seq-like methodology yielded a nearly identical splicing map for exons directly bound by NOVA¹². Importantly, these maps were able to predict novel NOVA-dependent regulated exons, with 30 of 51 predicted alternative exons showing NOVA-dependent regulation (all

thirty of which were correctly predicted as NOVA-enhancing or NOVA-repressing respectively) ⁶⁵.

Following this work, other analyses created such splicing maps for a variety of factors that revealed further insight into RNA regulatory processes. Some factors, such as RBFOX2, show similar characteristics to the maps generated for NOVA: RBFOX2 binds to a specific UGCAUG element which is associated with exon silencing when present in 3' end of the upstream intron, but associated with enhanced exon inclusion if located near the 5' splice site ¹³. In contrast, analysis of HNRNPC *in vivo* binding by iCLIP indicated enrichment near both the 5' and 3' ends of hnRNP C-silenced exons. In this case, further analysis focusing on only those loci with 160–170 nt of intervening sequence between two cross-linked nucleotides (which is indicative of full hnRNP particle formation) revealed further subtlety; full hnRNP particles were associated with exon inclusion in the upstream intron, but with silencing when bound to the exon itself ⁹. RNA splicing maps have now been generated for a variety of RNA binding proteins, revealing interesting similarities and differences in their effect at various positions along an alternative exon and its flanking introns ⁶⁶.

However, this simple model does not suffice for all factors. Analysis of *in vivo* targets of ALS-linked RNA binding protein FUS/TLS (fused in sarcoma/translocated in liposarcoma) revealed a unexpectedly broad pattern of binding. FUS binding was highly enriched for binding within long introns, where it exhibited a characteristic 'sawtooth-like' pattern (with highest association observed at the 5' splice site but decreasing throughout the intron) suggestive of co-transcriptional deposition. For FUS, intronic length proved to be one of the key features of whether a bound target would be differentially expressed upon FUS knockdown, suggesting that functional splicing maps for some RBPs may require a more complex model than simply localization of binding within the intron-alternative exon-intron cassette ⁶⁷.

The generation of RNA splicing maps has provided a useful analysis method to understand the roles of splicing regulators. The same type of analysis, incorporating either RNA half-life or ribosome occupancy, can similarly provide insight into the location-dependent roles an RBP plays in regulating other aspects of RNA processing. For example, incorporation of CLIP-seq, ribosome footprinting, and RNA-seq data for NMD factor UPF1 revealed that in addition to the traditional features characteristic of RNAs targeted for degradation by NMD (long 3' UTRs, presence of a premature stop codon upstream of the terminal exon junction, and presence of an upstream open reading frame), identification of upstream open reading frames with significant ribosome footprinting signal provided improved ability to predict NMD targets ⁶⁸. Additionally, UPF1 binding was predictive of degradation regardless of 3'UTR length, suggesting that the correlation between 3'UTR length and NMD may be either upstream or independent of UPF1-mediated degradation ⁶⁸. Further analysis using both RIP- and CLIP-seq of UPF1 ATPase-mutants revealed that it was not UPF1 recruitment to specific NMD targets, but rather regulation of UPF1 dissociation from non-target mRNAs, that controls which mRNAs will be ultimately degraded ⁶⁹. Although profiling of translation is not yet as commonly performed as RNA quantification, the recent commercial availability of ribosome footprinting protocols should enable further efforts to link RBP regulatory activity to not only RNA expression, but translational levels as well.

Integration of multiple RBP datasets

Integrated analysis of multiple RNA binding proteins in parallel presents an opportunity to learn the regulatory networks for individual proteins, the interplay between co-regulating factors, and build more complete RNA regulatory networks to explain altered RNA processing in a cell-type or disease of interest. Many efforts along these lines have focused on coordinated studies of paralogous factors, which often show complex co-regulation of targets.

The degree of functional redundancy can vary greatly among RNA binding protein families. Identification of *in vivo* targets of the three FMR1 RNA binding protein family members (FMRP, FXR1, and FXR2) revealed that >95% of FXR1 and FXR2 binding sites co-localized with FMRP binding sites⁷⁰. Similarly, analysis of MBNL1 and MBNL2 targets in various tissues across human and mouse suggested redundant roles in splicing regulation^{71,72}. In contrast, target identification of other factors revealed striking differences between binding patterns. RNA binding proteins TDP-43 and FUS/TLS have both been shown to play independent significant roles in amyotrophic lateral sclerosis (ALS), as mutations in both have been proposed to cause a significant fraction of inherited ALS cases⁷³. Although yielding similar phenotypes, identification of *in vivo* targets (by CLIP-seq) as well as identification of differentially regulated targets (identified by knockdown of TDP-43 and FUS/TLS in *in vitro* differentiated human neurons) showed largely distinct sets of targets at both the gene expression and splicing level⁶⁷. However, by focusing on the small number of genes that were bound by both TDP-43 and FUS/TLS, it was found that a small number of transcripts encoding genes with essential functions in neurons were downregulated upon knockdown of either TDP-43 or FUS/TLS, suggesting potential co-regulation of a small but biologically relevant subset of targets⁶⁷.

Intriguingly, other work has shown that binding interactions themselves can be dependent upon the combinatorial effect of other RNA binding proteins. Analysis of SRSF1 binding sites in cells in which SRSF2 is depleted by siRNA suggested that SRSF1 and SRSF2 compete for binding at similar sites, as SRSF1 binding was enriched at strong SRSF2 binding sites upon SRSF2 depletion⁷⁴. However, SRSF2 depletion led to reduction of SRSF1 binding at other sites, suggesting that at other (typically weaker) binding sites SRSF1 and SRSF2 binding is coupled through an unknown mechanism⁷⁴.

The emergence of high-throughput target identification methods enables large-scale cross-protein comparisons. The first such effort in *S. cerevisiae* profiled *in vivo* targets for 40 RNA binding proteins, including a number of proteins not previously annotated as interacting with RNA. The generation of these datasets using identical cells and methodology enabled cross-protein analyses, identifying novel co-binding between proteins and the first semi-global look at the combinatorial binding of dozens of RNA binding proteins in parallel⁷⁵. Similar efforts to profile RNA binding protein targets for many proteins in parallel have identified targets of four hnRNP proteins (hrp36, hrp38, hrp40, and hrp48) in *Drosophila*, 12 assorted RNA binding proteins (PUM2, QKI, IGF2BP1-3, AGO1-4, and TNRC6A-C) in human HEK293 cells, six HNRNP family members (HNRNPA1, HNRNPA2/B1, HNRNPF, HNRNPH1, HNRNPM, and HNRNPU) in human 293T cells, and four eIF3 translation

initiation complex members (eIF3a, eIF3b, eIF3d, and eIF3G) in human 293T cells^{10,29,76,77}. In each case, the generation of target information in a standardized manner in the same cell-type enabled cross-dataset comparisons that revealed unexpected complexity in factor co-association and co-regulation of targets. These types of analyses represent both opportunities and challenges: the opportunity to obtain the first truly global views of RNA processing regulation, but the challenge of developing computational methods to integrate not only multiple experiments for an individual RBP, but to also analyze dozens (or hundreds) of RBPs in parallel. This includes the significant complexity inherent in analyzing RNA binding proteins that are often alternatively spliced or otherwise regulated at the RNA level themselves, leading to substantial cross-regulation among RBPs²⁹. Although these initial analyses have largely focused on individual RBP regulation with a brief consideration for combinatorial regulation, the next stage of these efforts will likely turn to machine learning algorithms to help to fully understand the fully complexity of the human RNA processing regulatory network.

In addition to providing RBP-specific insights, integrative analysis may reveal previously unknown properties of RNA processing and regulation. An interesting parallel can be seen for studies of transcription factor targets: although identification of *in vivo* binding sites for many individual transcription factors (by chromatin immunoprecipitation followed by high-throughput sequencing, or “ChIP-seq”) has yielded significant insights into the biological roles of those individual factors, large-scale ChIP-seq efforts performed by individual labs as well as the ENCODE project (in human) and modENCODE project (in *Drosophila* and *C. elegans*) revealed novel general properties of transcriptional regulation^{78,79}. In particular, these analyses led to the characterization of HOT regions (loci bound by the majority of assayed factors) as a novel regulatory mechanism for regulation of essential housekeeping genes^{78–80}, and led to the development of models to predict target gene expression based on a variety of features (including both sequence and binding information)⁸¹. These results strongly suggest that in addition to RBP-specific insights, it remains possible that these large-scale efforts to profile RNA regulatory networks may also reveal completely novel principles in regulation of RNA processing.

Conclusion

Over the coming years, our knowledge of RBP targets will continue to rapidly expand. In addition to the efforts of individual labs, which have identified targets for dozens of RBPs in various cell lines and tissues and will continue to deeply explore RNA processing regulation, the ENCODE project has now added an effort to profile the targets of ~250 annotated or predicted RNA binding proteins in two standard laboratory human cell lines. With these large-scale data generation efforts come significant computational challenges in automatically processing these datasets, subjecting them to automated quality control procedures, analyzing thousands of datasets in a rapid yet accurate manner, visualizing these analyses, and (most importantly) incorporating this information in order to infer novel aspects of RNA biology. Although many challenges remain, the work of numerous groups over the past few years have revealed significant insights into how RNA binding proteins act to regulate RNA processing across various tissues and cell-types, and how alteration of these regulatory activities can lead to disease. The further development of computational tools in

the coming years will allow a rapid expansion of this research, enabling a global picture of RNA processing that should lead to further insights into the roles that RNA processing plays in defining gene expression in humans.

Acknowledgments

This work was partially supported by grants from the National Institute of Health (HG007005, HG004659 and NS075449) to G.W.Y. G.W.Y. is an Alfred P. Sloan Research Fellow. E.L.V.N. is a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13).

References

1. Muller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature reviews. Genetics*. 2013; 14:275–287. DOI: 10.1038/nrg3434
2. Scherrer T, Mittal N, Janga SC, Gerber AP. A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PloS one*. 2010; 5:e15499. [PubMed: 21124907]
3. Tsvetanova NG, Klass DM, Salzman J, Brown PO. Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. *PloS one*. 2010; 5
4. Castello A, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*. 2012; 149:1393–1406. DOI: 10.1016/j.cell.2012.04.031 [PubMed: 22658674]
5. Kwon SC, et al. The RNA-binding protein repertoire of embryonic stem cells. *Nature structural & molecular biology*. 2013
6. Paige JS, Wu KY, Jaffrey SR. RNA mimics of green fluorescent protein. *Science*. 2011; 333:642–646. DOI: 10.1126/science.1207339 [PubMed: 21798953]
7. Nicol C, et al. An RNA aptamer provides a novel approach for the induction of apoptosis by targeting the HPV16 E7 oncoprotein. *PloS one*. 2013; 8:e64781. [PubMed: 23738000]
8. Shui B, et al. RNA aptamers that functionally interact with green fluorescent protein and its derivatives. *Nucleic acids research*. 2012; 40:e39. [PubMed: 22189104]
9. König J, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*. 2010; 17:909–915. DOI: 10.1038/nsmb.1838
10. Hafner M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*. 2010; 141:129–141. DOI: 10.1016/j.cell.2010.03.009 [PubMed: 20371350]
11. Ule J, et al. CLIP identifies Nova-regulated RNA networks in the brain. *Science*. 2003; 302:1212–1215. DOI: 10.1126/science.1090095 [PubMed: 14615540]
12. Licatalosi DD, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*. 2008; 456:464–469. DOI: 10.1038/nature07488 [PubMed: 18978773]
13. Yeo GW, et al. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*. 2009; 16:130–137. DOI: 10.1038/nsmb.1545
14. Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature biotechnology*. 2011; 29:607–614. DOI: 10.1038/nbt.1873
15. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007; 316:1497–1502. DOI: 10.1126/science.1141319 [PubMed: 17540862]
16. Valouev A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*. 2008; 5:829–834. DOI: 10.1038/nmeth.1246 [PubMed: 19160518]
17. Cooper TA. Use of minigene systems to dissect alternative splicing elements. *Methods*. 2005; 37:331–340. DOI: 10.1016/j.ymeth.2005.07.015 [PubMed: 16314262]
18. Lambeth LS, Smith CA. Short hairpin RNA-mediated gene silencing. *Methods in molecular biology*. 2013; 942:205–232. DOI: 10.1007/978-1-62703-119-6_12 [PubMed: 23027054]
19. Rao DD, Senzer N, Cleary MA, Nemunaitis J. Comparative assessment of siRNA and shRNA off target effects: what is slowing clinical development. *Cancer gene therapy*. 2009; 16:807–809. DOI: 10.1038/cgt.2009.53 [PubMed: 19713999]

20. Ho TH, et al. Muscleblind proteins regulate alternative splicing. *The EMBO journal*. 2004; 23:3103–3112. DOI: 10.1038/sj.emboj.7600300 [PubMed: 15257297]
21. Tang ZZ, Zheng S, Nikolic J, Black DL. Developmental control of CaV1.2 L-type calcium channel splicing by Fox proteins. *Molecular and cellular biology*. 2009; 29:4757–4765. DOI: 10.1128/MCB.00608-09 [PubMed: 19564422]
22. Gehman LT, et al. The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nature genetics*. 2011; 43:706–711. DOI: 10.1038/ng.841 [PubMed: 21623373]
23. McPherson LA, Weigel RJ. AP2alpha and AP2gamma: a comparison of binding site specificity and trans-activation of the estrogen receptor promoter and single site promoter constructs. *Nucleic acids research*. 1999; 27:4040–4049. [PubMed: 10497269]
24. Sarkar B, Lu JY, Schneider RJ. Nuclear import and export functions in the different isoforms of the AUF1/heterogeneous nuclear ribonucleoprotein protein family. *The Journal of biological chemistry*. 2003; 278:20700–20707. DOI: 10.1074/jbc.M301176200 [PubMed: 12668672]
25. Johnson JM, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*. 2003; 302:2141–2144. DOI: 10.1126/science.1090100 [PubMed: 14684825]
26. Shen S, Warzecha CC, Carstens RP, Xing Y. MADS+: discovery of differential splicing events from Affymetrix exon junction array data. *Bioinformatics*. 2010; 26:268–269. DOI: 10.1093/bioinformatics/btp643 [PubMed: 19933160]
27. Shai O, Morris QD, Blencowe BJ, Frey BJ. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics*. 2006; 22:606–613. DOI: 10.1093/bioinformatics/btk028 [PubMed: 16403798]
28. Yeo GW, Van Nostrand EL, Liang TY. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS genetics*. 2007; 3:e85. [PubMed: 17530930]
29. Huelga SC, et al. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell reports*. 2012; 1:167–178. DOI: 10.1016/j.celrep.2012.02.001 [PubMed: 22574288]
30. Du H, et al. Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. *Nature structural & molecular biology*. 2010; 17:187–193. DOI: 10.1038/nsmb.1720
31. Chen, Lin. *Handbook of Statistical Bioinformatics*. Schölkopf, Bernhard, Lu, Henry, Horng-Shing, Zhao, Hongyu, editors. Springer; Berlin Heidelberg: 2011. p. 31-53.
32. Parkhomchuk D, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic acids research*. 2009; 37:e123. [PubMed: 19620212]
33. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29:15–21. DOI: 10.1093/bioinformatics/bts635 [PubMed: 23104886]
34. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*. 2009; 10:R25. [PubMed: 19261174]
35. Trapnell C, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*. 2013; 31:46–53. DOI: 10.1038/nbt.2450
36. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456:470–476. DOI: 10.1038/nature07509 [PubMed: 18978772]
37. Katz Y, Wang ET, Airolidi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*. 2010; 7:1009–1015. DOI: 10.1038/nmeth.1528 [PubMed: 21057496]
38. Shepard PJ, et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *Rna*. 2011; 17:761–772. DOI: 10.1261/rna.2581711 [PubMed: 21343387]
39. Derti A, et al. A quantitative atlas of polyadenylation in five mammals. *Genome research*. 2012; 22:1173–1183. DOI: 10.1101/gr.132563.111 [PubMed: 22454233]
40. Hutvagner G, Simard MJ. Argonaute proteins: key players in RNA silencing. *Nature reviews. Molecular cell biology*. 2008; 9:22–32. DOI: 10.1038/nrm2321 [PubMed: 18073770]
41. Houseley J, Tollervey D. The many pathways of RNA degradation. *Cell*. 2009; 136:763–776. DOI: 10.1016/j.cell.2009.01.019 [PubMed: 19239894]

42. Lal A, et al. Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *The EMBO journal*. 2004; 23:3092–3102. DOI: 10.1038/sj.emboj.7600305 [PubMed: 15257295]
43. Tani H, et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome research*. 2012; 22:947–956. DOI: 10.1101/gr.130559.111 [PubMed: 22369889]
44. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009; 324:218–223. DOI: 10.1126/science.1168978 [PubMed: 19213877]
45. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. 2011; 147:789–802. DOI: 10.1016/j.cell.2011.10.002 [PubMed: 22056041]
46. Schwanhaussner B, et al. Global quantification of mammalian gene expression control. *Nature*. 2011; 473:337–342. DOI: 10.1038/nature10098 [PubMed: 21593866]
47. Gry M, et al. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC genomics*. 2009; 10:365. [PubMed: 19660143]
48. Kondrashov N, et al. Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell*. 2011; 145:383–397. DOI: 10.1016/j.cell.2011.03.028 [PubMed: 21529712]
49. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079. DOI: 10.1093/bioinformatics/btp352 [PubMed: 19505943]
50. <<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>>
51. DeLuca DS, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*. 2012; 28:1530–1532. DOI: 10.1093/bioinformatics/bts196 [PubMed: 22539670]
52. Robinson JT, et al. Integrative genomics viewer. *Nature biotechnology*. 2011; 29:24–26. DOI: 10.1038/nbt.1754
53. Kent WJ, et al. The human genome browser at UCSC. *Genome research*. 2002; 12:996–1006. Article published online before print in May 2002. DOI: 10.1101/gr.229102 [PubMed: 12045153]
54. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science*. 2002; 297:1007–1013. DOI: 10.1126/science.1073774 [PubMed: 12114529]
55. Dror G, Sorek R, Shamir R. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*. 2005; 21:897–901. DOI: 10.1093/bioinformatics/bti132 [PubMed: 15531599]
56. Sorek R, et al. A non-EST-based method for exon-skipping prediction. *Genome research*. 2004; 14:1617–1623. DOI: 10.1101/gr.2572604 [PubMed: 15289480]
57. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:2850–2855. DOI: 10.1073/pnas.0409742102 [PubMed: 15708978]
58. Barash Y, et al. Deciphering the splicing code. *Nature*. 2010; 465:53–59. DOI: 10.1038/nature09000 [PubMed: 20445623]
59. Xiong HY, et al. RNA Splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*. 2015; 347:1254806. [PubMed: 25525159]
60. Faber K, Glatting KH, Mueller PJ, Risch A, Hotz-Wagenblatt A. Genome-wide prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASites. *BMC bioinformatics*. 2011; 12(Suppl 4):S2.
61. Yang YY, Yin GL, Darnell RB. The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. *Proceedings of the National Academy of Sciences of the United States of America*. 1998; 95:13254–13259. [PubMed: 9789075]
62. Jensen KB, et al. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*. 2000; 25:359–371. [PubMed: 10719891]
63. Buckanovich RJ, Darnell RB. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Molecular and cellular biology*. 1997; 17:3194–3201. [PubMed: 9154818]

64. Lewis HA, et al. Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell*. 2000; 100:323–332. [PubMed: 10676814]
65. Ule J, et al. An RNA map predicting Nova-dependent splicing regulation. *Nature*. 2006; 444:580–586. DOI: 10.1038/nature05304 [PubMed: 17065982]
66. Witten JT, Ule J. Understanding splicing regulation through RNA splicing maps. *Trends in genetics : TIG*. 2011; 27:89–97. DOI: 10.1016/j.tig.2010.12.001 [PubMed: 21232811]
67. Lagier-Tourenne C, et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nature neuroscience*. 2012; 15:1488–1497. DOI: 10.1038/nn.3230 [PubMed: 23023293]
68. Hurt JA, Robertson AD, Burge CB. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome research*. 2013; 23:1636–1650. DOI: 10.1101/gr.157354.113 [PubMed: 23766421]
69. Lee SR, Pratt GA, Martinez FJ, Yeo GW, Lykke-Andersen J. Target Discrimination in Nonsense-Mediated mRNA Decay Requires Upf1 ATPase Activity. *Molecular cell*. 2015; 59:413–425. DOI: 10.1016/j.molcel.2015.06.036 [PubMed: 26253027]
70. Ascano M Jr, et al. FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*. 2012; 492:382–386. DOI: 10.1038/nature11737 [PubMed: 23235829]
71. Wang ET, et al. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*. 2012; 150:710–724. DOI: 10.1016/j.cell.2012.06.041 [PubMed: 22901804]
72. Han H, et al. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*. 2013; 498:241–245. DOI: 10.1038/nature12270 [PubMed: 23739326]
73. Lagier-Tourenne C, Polymenidou M, Cleveland DW. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Human molecular genetics*. 2010; 19:R46–64. DOI: 10.1093/hmg/ddq137 [PubMed: 20400460]
74. Pandit S, et al. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Molecular cell*. 2013; 50:223–235. DOI: 10.1016/j.molcel.2013.03.001 [PubMed: 23562324]
75. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS biology*. 2008; 6:e255. [PubMed: 18959479]
76. Blanchette M, et al. Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the Drosophila hnRNP A/B family members. *Molecular cell*. 2009; 33:438–449. DOI: 10.1016/j.molcel.2009.01.022 [PubMed: 19250905]
77. Lee AS, Kranzusch PJ, Cate JH. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature*. 2015; 522:111–114. DOI: 10.1038/nature14267 [PubMed: 25849773]
78. Gerstein MB, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*. 2010; 330:1775–1787. DOI: 10.1126/science.1196914 [PubMed: 21177976]
79. mod EC, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*. 2010; 330:1787–1797. DOI: 10.1126/science.1198374 [PubMed: 21177974]
80. Gerstein MB, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012; 489:91–100. DOI: 10.1038/nature11245 [PubMed: 22955619]
81. Marbach D, et al. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome research*. 2012; 22:1334–1349. DOI: 10.1101/gr.127191.111 [PubMed: 22456606]

Figure 1. Overview of methods to identify RNA binding protein regulation of various RNA processing steps

(A) RNA molecules bound by RNA binding proteins can be identified by a variety of related approaches generally termed CLIP-seq. Cell lysates are crosslinked (often with UV, to selectively cross-link protein-RNA interactions), and limited RNase digestion is performed to generate small protein-bound RNA fragments. A RBP-specific antibody is then used to immunoprecipitate the protein of interest, along with associated RNA. After protein digestion, linkers are ligated to the 5' and 3' end of RNA fragments, which are then used to reverse transcribe and PCR amplify a DNA library for high throughput sequencing. (B) mRNAs being actively translated are identified using ribosome protection assays. Using the ART-seq kit (available from Epicentre), ribosome protected fragments are isolated by nuclease digestion followed by monosome purification. RNA is then purified, annealed to adapter sequences, and amplified for high-throughput sequencing. (C) Transcriptome-wide RNA expression and alternative splicing can be queried by RNA-seq (left) or microarray (right). For RNA-seq (left), ribosomal RNA is depleted using alternative methods, RNA is fragmented and reverse transcribed into cDNA, sequencing adapters are added, and then amplified for high-throughput sequencing. Stranded RNA-seq libraries can be generated by incorporating dUTP during second strand cDNA synthesis, followed by cleavage by UDG enzymes. For microarrays (right), RNA is converted to cDNA, amplified, fragmented, and fluorescently labeled. Labeled DNA is then hybridized to the microarray, and expression is read out by fluorescence intensity. (D) Other aspects of RNA processing can be studied

using specialized protocols, such as BRIC-seq-based quantitation of RNA half-lives. After growth in media containing 5'-bromo-uridine (BrU), fresh media is added to terminate labeling of newly synthesized RNAs with BrU. Pull-down of remaining RNA with an anti-BrU antibody at various timepoints, followed by library preparation using standard RNA-seq protocols, can then yield quantitative estimates of RNA stability.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2. Identification of direct RNA binding protein targets by CLIP-seq

(a) Overview of CLIP-seq procedure (repeated from Figure 1). (B) Several variants of the general CLIP-seq method (left) have been developed to optimize various steps. iCLIP (center) uses a single linker oligonucleotide coupled with a circular ligase enzyme to improve linker ligation efficiency⁹. This linker also adds a barcode as well as five random nucleotides to each sequencing read, allowing one to disambiguate PCR amplification artifacts from truly independent RNA fragments, increasing the dynamic range of sequencing read depth. In PAR-CLIP (right), photo-reactive ribonucleotide analogs are

incorporated into transcribed RNA in order improve crosslinking efficiency¹⁰. In addition, the cross-linked uridine will be complemented by a guanine base in the reverse transcribed cDNA during library preparation step. This characteristic U to C mutation in the center of a binding site can be used to identify binding at single-nucleotide resolution. Single-nucleotide resolution binding sites can also be computationally identified from normal CLIP-seq (where reverse transcription often skips the cross-linked nucleotide, leading to single-base deletions (green triangle) in the sequencing reads), or iCLIP (where reverse transcription often terminates at the crosslinked base pair, leading to read pileups at the crosslink site).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3. Analysis of alternative RNA processing by microarray

(A) Overview of RNA expression quantification by microarray (repeated from Figure 1). (B) Quantitation of alternative splicing events by microarray requires probes designed to target specific regions associated with the event. For a cassette exon, traditional exon microarrays will have probes assaying the desired alternatively spliced (AS) exon (purple) as well as flanking (UP and DN) constitutive exons (blue). More recent microarrays have incorporated additional probes assaying splice junctions (UP jxn and DN jxn) that quantify exon inclusion (red) or the alternatively spliced (AS) exclusion junction (green). (C) Splicing-sensitive microarrays identify alternative splicing of *SLK* exon $\times$ during neuronal differentiation. (Left) Although flanking exons show relatively similar signal intensity between control and knockdown samples, the three probe sets querying exon inclusion show ~8-fold higher signal in iPS cells as compared to differentiated neurons, whereas the probeset measuring

the junction created by exon exclusion increases ~4-fold. (F.J.M. & G.Y. *unpublished data*). (center) As a first step to identify significantly altered splicing events, change in probe intensity for splicing-sensitive probes are normalized for gene expression changes by regression of all other constitutive exon probes covering the SLK gene. (right) Various algorithms have been described to convert the characteristic opposing changes in UPjxn/DNjxn and ASjxn probe intensity into an estimate of PSI value. These PSI values can then be validated by RT-PCR.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 4. Identification of altered RNA processing by RNA-seq

A. Schematic representation of sequencing reads enabling quantitation and identification of an alternative cassette exon. Reads mapping to the upstream and downstream flanking exons (black) can be used to estimate the overall level of transcript abundance but are not informative for splicing of the internal exon. Reads mapping to the alternative exon (purple) or to the upstream or downstream splice junctions (red) provide evidence for exon inclusion, whereas reads mapping to the junction between the flanking exons (blue) provide evidence for exon exclusion. (B) Similar schematic for alternative polyadenylation site usage. Identification of alternative polyadenylation sites is more complex, as there are reads that uniquely map to the longer isoform (red), but all other reads are expected to map to both isoforms (blue). (C-D) Read density profile for exon 4 of *EEF1B2* in three RNA-seq experiments, with height of the blue histogram indicating the number of reads that map with the 5' end at the given base. The percent spliced in (Ψ) value is then defined as the number of reads supporting inclusion divided by the number supporting either inclusion or exclusion. Significance relative to a control sample can then be calculated by Fisher's Exact test (or the equivalent Chi-square test) on the number of inclusion- and exclusion-supporting reads.