Introduction

Polyadenylation of RNA is pervasive in the animal kingdom. In eukaryotes messenger RNA (mRNA) is polyadenylated at the 3' end as part of pre mRNA processing, and this poly(A) tail facilitates the transport and stability of the mRNA [**?**]. In bacteria, polyadenylation is known for the opposite effect; the addition of a poly(A) tail usually marks an RNA for degradation [**?**]. Recently, however, polyadenylation as a signal for degradation was also identified in the yeast nucleus [**?**] [] CIT. Here polyadenylation has two roles: signalling for degradation in the nucleus and for protection from degrdation in the cytoplasm Even more recently, polyadenylation that does not come from pre mRNA processing has been found in human too CIT and has also been linked to transcript degradation both in the nucleus and cytoplasm CIT. Recently, RNA-seq has been used to identify and study novel polyadenylation sites in human, yeast, and worm CIT CIT CIT. Since standard RNA-seq protocols often give poor coverge of 3' ends, specialiced proceedures were developed in CIT and CIT to obtain a high coverage of polyadenylation reads. While polyadenylated RNAs exists in both the nucleus and cytoplasm, the studies so far have investigated only whole-cell extracts. In RNA-seq, whole cell extracts are enriched for RNA from the cytoplasm, simply because the RNA in the cytoplasm dominates the sample by volume. This gives a lower resolution of polyadenylation in the nucleus. Here we investigate polyadenylation in human cell lines from the ENCODE project in the the nucleic and cytoplasmic compartments. We identify an enrichment of polyadenylation signals in introns and speculate that these may be by-products of the degradation of the introns in the nucleoplasm. [**?**], []<++> [**?**] [**?**]q

Methods

We mapped the RNA-seq data to the human genome, screened and trimmed the unmapped reads for leading T or trailing As, and remapped the trimmed reads to the genome. A trimmed read was accepted as a read that stemmed from a polyadenylation event if there was no poly(A/T) stretch on the region of the genome which corresponds to the trimmed part of the read.

The reads were also screened against split-mapped reads to make sure that the poly(A) reads were not just exon-exon junction reads over an A/T rich region.

The poly(A) reads were clustered together to form poly(A) clusters in the same manner as []. Using the terminology from CIT, we refer to the poly(A) reads as polyadenylation tags (PAT), and the clusters of PATs as polyadenylation tag clusters (PAC). After clustering, we searched the downstream sequence of each cluster for the polyadenylation signal (PAS). We also looked in a window of 30 nucleotides around the polyadenylation site for an annotated poly(A) site. The set of annotated poly(A) sites was obtained by merging the poly(A) site annotation of GENCODE with the pA db[] to obtain 43000 annotated polyadenylation sites.

5UTR, CDS, and 3UTR regions were extracted from the GENCODE annotation version 7 []. Regions were extracted in a non-overlapping fashion with the following precedence: exons $>$ introns, 3UTR $>$ 5UTR $>$ CDS.

To map the reads to the genome we used the gem-mapper CIT.

Results

Polyadenylation in the 3' UTR

Although the RNA-seq protocol used for the data in this study is optimized for 3'UTR ends we find abundant polyadenylation signals in the 3UTRs of human mRNAs. In the cytocolic compartments of K562, GM.., and HeLa we find XXX, XXX, and XXX PAC. To be counted, a PAC must have 2 or more PAT or land at an annotated polyadenylation site. Of XXX PAC, YYY have only 1 PET. Of these YYY PEC, however, XYX % land at an annotated PAC site. Figure XX shows the relationship between increasing the number of reads and the number of poly(A) sites discovered in the 3UTR. GGG % of the PAC in the 3UTRs have one of the 8 PAS listed in CIT.

Nucleoplasm and cytoplasm vs. whole cell

Figure XXX shows the number of poly(A) sites in different genomic regions for nucleoplasm, cytoplasm, and whole cell for K562. As can be seen, the whole cell extract is most similar to the cytosolic extract. Most prominently, one sees that the introns are poorly represented in the whole cell extract.

Evidence for polyadenylation-events in introns

There is an enrichment of poly(A) reads in introns. In contrast to the poly(A) reads in the 3UTR, few of these reads have the PAS signal (XX %), which supports the notion that these poly(A) signals are of a different nature than the ones found at the 3' of mRNAs. By contrast, XYX % of the the PAC in the 3UTR with only 1 poly(A) read have downstream PAS.

TODO: compare the poly(A) reads without PAS in introns and 5'UTR exons. Are there more or less in introns normalized for sequence length?

Discussion

It has recently become clear that human RNA is adenylated, likely in a transient, degradation-linked manner CIT. If this adenylation has the same purpose as in yeast and bacteria, adenylated tails may functions as docking stations for the exosome, something which might be especially important if the RNA is structured at the 3' end, preventing access to exonucleases CIT.

Why should we believe these results? Few because the process is transient. The polyadenylated ones are quickly degraded, not stable like the other polyAz. What do the polyadenylation events mean?