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 [1]. In bacteria, polyadenylation is known for the opposite effect; the addition of a poly(A) tail usually marks an RNA for degradation [8]. Recently, however, polyadenylation as a signal for degradation was also identified in the yeast nucleus [4] [12]. 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 [6] and has also been linked to transcript degradation both in the nucleus and cytoplasm [10].

In several studies RNA-seq has been used to identify and study novel polyadenylation sites in human, yeast, and worm [3] [9] [7]. Since standard RNA-seq protocols often give poor coverge of 3' ends, specialiced proceedures were developed [9] [3] [2] to obtain a higher coverage of polyadenylation reads. These studies have used whole cell extracts as substrates for the sequencing. Since there is comparably less RNA in the nucleus than in the cytoplasm, using whole cell extracts gives a comparably lower resolution to the polyadenylation happening in the nucleus. Further, the cytoplasm and the nucleus harbor different types of RNA. Connected to chromatin is nascent RNA, but unprocessed and mid-processing (splicing and polyadenylation); in the nucleoplasm are both successfully produces mRNA ncRNA, snoRNA etc, as well as mis-processed RNA that is being degraded, and perhaps the largest RNA group in the nucleoplasm are the introns that are being degraded.

Here we investigate polyadenylation in human cell lines from the ENCODE project in the the nucleic and cytoplasmic compartments separately. We find extensive evidence of polyadenylation in 3UTRs both in the cytoplasm and the nucleus, and we also find non-3UTR linked polyadenylation among other places in introns. We speculate that the non-3UTR-linked polyadenylation signals result from degradation-in-process.

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 cleavage sites were clustered together to form poly(A) clusters (PAC) as in [11]. We consider those PAC that have two or more cleavage sites supporting them or fall at annotated cleavage sites. 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 polyAdb [5] to

obtain 43000 annotated polyadenylation sites.

The 5UTR, CDS, and 3UTR regions (and their introns) 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. (http://gemlibrary.sourceforge.net/)

Results

Identifying known polyadenylation cytoplasmic sites in the 3' UTR

To verify that our method is able to find known polyadenylation sites, we investigated the PAC that land in annotated 3UTR regions. Combining the cleavage sites for HeLA, K562, and GM12878, we obtained XXX clusters, of which XXX land at annotated poly(A) sites, and XXY have PAS. Thus in addition to identifying XXX annotated sites, we also identify XYX potentially novel sites, YYX % of which have a PAS. In figure 1a can be seen the cumulative number of poly(A) sites obtained, as well as the relationship between the RPKM of annotated 3UTRs with the number of poly(A) sites identified. As can be seen, due to not using a 3'-optimized RNA-Seq protocol, the highly expressed transcripts are most likely to have their poly(A) sites identified. However, the ones that are found are of high specificity, since most land at annotated poly(A) sites, and of those that do not, a high percentage have the PAS signal downstream the putative cleavage site.

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.

Difference in poly(A) tail nucleotide composition and length between introns and 3UTRs

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.

Further, the tail-lengths of the poly(A) reads is different. Also comment upon the nr and distribution of A-runs and of T-runs. Seems the average length is shorter in the 3UTRs!

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?

Using standard RNAseq protocol for polyA

Why do people want to find poly(A) for their dataset? Instead of running a separate technique, they can apply their standard protocol. 75 vs 150 etc. compare with that 36 read length one, the oldest. they got very few. we get much more. how many with 150?

References

[1] Diana F. Colgan and James L. Manley. Mechanism and regulation of mRNA polyadenylation. *Genes & Development*, 11(21):2755 –2766, November 1997.

[2] Kristi Fox-Walsh, Jeremy Davis-Turak, Yu Zhou, Hairi Li, and Xiang-Dong Fu. A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation. *Genomics*, In Press, Corrected Proof.

[3] Yonggui Fu, Yu Sun, Yuxin Li, Jie Li, Xingqiang Rao, Chong Chen, and Anlong Xu. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Research*, 21(5):741 –747, May 2011.

[4] John LaCava, Jonathan Houseley, Cosmin Saveanu, Elisabeth Petfalski, Elizabeth Thompson, Alain Jacquier, and David Tollervey. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell*, 121(5):713–724, June 2005.

[5] Ju Youn Lee, Ijen Yeh, Ji Yeon Park, and Bin Tian. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Research*, 35(Database issue):D165–D168, January 2007. PMID: 17202160 PMCID: 1899096.

[6] Carol S. Lutz and Alexandra Moreira. Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression. *Wiley Interdisciplinary Reviews: RNA*, 2(1):22–31, January 2011.

[7] Marco Mangone, Arun Prasad Manoharan, Danielle Thierry-Mieg, Jean Thierry-Mieg, Ting Han, Sebastian D. Mackowiak, Emily Mis, Charles Zegar, Michelle R. Gutwein, Vishal Khivansara, Oliver Attie, Kevin Chen, Kourosh Salehi-Ashtiani, Marc Vidal, Timothy T. Harkins, Pascal Bouffard, Yutaka Suzuki, Sumio Sugano, Yuji Kohara, Nikolaus Rajewsky, Fabio Piano, Kristin C. Gunsalus, and John K. Kim. The landscape of c. elegans 3′UTRs. *Science*, 329(5990):432 –435, July 2010.

[8] Bijoy K Mohanty and Sidney R Kushner. Bacterial/archaeal/organellar polyadenylation. *Wiley Interdisciplinary Reviews: RNA*, 2(2):256–276, March 2011.

[9] Fatih Ozsolak, Philipp Kapranov, Sylvain Foissac, Sang Woo Kim, Elane Fishilevich, A. Paula Monaghan, Bino John, and Patrice M. Milos. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143(6):1018–1029, December 2010.

[10] Shimyn Slomovic, Ella Fremder, Raymond H. G. Staals, Ger J. M. Pruijn, and Gadi Schuster. Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proceedings of the National Academy of Sciences*, 107(16):7407 –7412, April 2010.

[11] Bin Tian, Jun Hu, Haibo Zhang, and Carol S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201 –212, January 2005.

[12] Françoise Wyers, Mathieu Rougemaille, Gwenaël Badis, Jean-Claude Rousselle, Marie-Elisabeth Dufour, Jocelyne Boulay, Béatrice Régnault, Frédéric Devaux, Abdelkader Namane, Bertrand Séraphin, Domenico Libri, and Alain Jacquier. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new Poly(A) polymerase. *Cell*, 121(5):725–737, June 2005.