# Two types of polyadenlyation of human RNA from RNA-seq of cellular compartments

JÃ¸rgen Skancke

August 19, 2011

## Introduction

Polyadenylation of RNA is pervasive in the animal kingdom. In bacteria, polyadenylation marks RNA for degradation, while in eukaryotes mRNA are polyadenylation as part of pre-mRNA processing, along with splicing and 5' capping.

Polyadenylation of eukaryotic messenger RNA (mRNA) happens co-transcriptionally and is linked to transcription termination []. Both the site of polyadenylation on pre-mRNA and the poly(A) tail itself affect the regulation of transport and degradation [], []. In particular, in the cytoplasm the poly(A) tail protects against degradation. A much commented difference between eukaryotes and bacteria is that while polyadenylation protects against degradation in eukaryotes, it facilitates degradation in bacteria. However, recently reports have come of polyadenylation events in eukaryotes that are linked to the degradation of RNA. Initially these reports were from yeast, [], but recently reports have come of degrdation-related polyadenylation in humans too []. In yeast the non-canonical polyadenylation has been liked to the degradation of several types of transcripts, while in human they have so far only been reported for ribosomal RNA (rRNA).

We used RNA-seq data from the ENCODE project to do a genome-wide profiling of polyadenylation sites in the chromatin, nucleoplasmic, nuclear, and cytoplasmic compartments.

Here we show evidence for non-canonical degradation-linked polyadenylation of human RNA. Specifically, we show evidence of polyadenylation of introns in the nucleus. Intronic polyadenylation would not be linked to transcript termination, but rather to the degradation of these introns.

## Methods

We mapped the RNA-seq data to the human genome and screened the unmapped reads for leading T or trailing As, which were subsequently trimmed. The reads were then remapped to the genome. The 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 screend 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 []<++>. 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.

Genomic 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.

## Results

### Datasets

We investigated the HeLa, K563, and GM12872 cell lines from the ENCODE project []. HeLa and GM12878 have RNA-seq data for the nuclear and cytosolic compartments, both for poly(A)+ and poly(A)- RNA. For K562 there is also chromatin and nucleoplasm data available, although for these compartments the RNA has not been split into poly(A)+ and poly(A)- fractions.

### Canonical polyadenylation

To confirm that we can pick up the transcription-termination related polyadenylation (henceforth called canonical polyadenylation), we intersected the polyadenylation sites we find with different genomic regions. The result can be seen in figure 1. As can be seen, for the cytoplasmic compartments, there is a strong enrichment of poly(A) clusters in the 3UTRs, a XXX fold increase compared to the 5UTR and CDS combined. Of those poly(A) clusters that land in the 3UTR, XXX% land at annoated poly(A) sites and XXX% have a PAS signal associated with them. This confirms that the method of obtaining poly(A) reads manages to cover annoated poly(A) sites. Since the RNA-seq method is not optimized for 3' ends, we do not cover all poly(A) sites. The higher the RPKM , the more frequent are poly(A) reads found for a given gene (data not shown). Regardless, a good portion of poly(A) sites, both novel and annotated, are found with these data. See Figure 2 for an intersection between the total number of poly(A) sites found in the cytoplasmic regions and the annotated poly(A) sites. See Table 1 for summary-statistics for the poly(A) clusters from the different compartments.