Polyadenylation in different cellular compartments

August 26, 2011

Introduction

In eukaryotes polyadenylation occurs after 3' cleavage as one of the last stages of pre-mRNA processing. The cleavage/polyadenylation process is triggered by the polyadenylation signal (PAS) which is typically found 10-30 nucleotides downsteam the cleavage site. More generally, adenylation of RNA is found in nearly all kingdoms of life, but plays opposite roles in bacteria and eukaryotes, where in bacteria adenylation promotes degradation of RNA, while in eukaryotes an RNA's poly(A) tail protects from degradation.

We investigated the RNA-seq data for 5 GENC ODE cell lines for signs of polyadenylation by trimming and remapping those reads that were originally unmappable and ended in a stretch of As or beginning with a stretch of Ts. We subsequently merged the polyadenylation sites into clusters in a similar way to Tian et. al [4]. In total, for all cell lines and compartments, we obtained 29860 putative polyadenylation sites. A putative polyadenylation site is here defined as a site at which two or more trimmed reads cluster together, or a site at which a read lands at a previously annotated polyadenylation site. Annotated polyadenylation sites were found by merging and clustered 43187 sites from the polyAdb [1] with 35791 sites from GENCODE to obtain a total of 50696 annotated polyA sites. 16657 of our polyA sites fall at annotated ones (81% or 13517 of these are in 3UTRs), leaving a total of 13203 putative novel polyadenylation sites in the genome (Fig 1

Polyadenylation in the nucleus and in the cytoplasm

The RNA-seq protocol for the Gingeras data is not optimized for 3' ends, therefore we expected to find most poly(A) sites for transcripts with high RPKM. To investigate this, we calculated the ratio of discovered to annotated poly(A) sites in 3UTRs that do not overlap any other genomic feature. Figure 2 shows the relationship between RPKM and poly(A) discovery ratio for annotated 3UTRs. As can be seen, there is a positive association between RPKM and poly(A) discovery ($r = 0.52$, $p < 10^{-10}$), however there is considerable variation even for high RPKM transcripts. The average number of poly(A) sites found per 3UTR was 1.7, and the average ratio of poly(A) sites to annotated was 0.9. That fewer than 1 per annotated site are found probably reflects both that the method is not exhaustive, and that not all annotated poly(A) sites are in use at a given time.

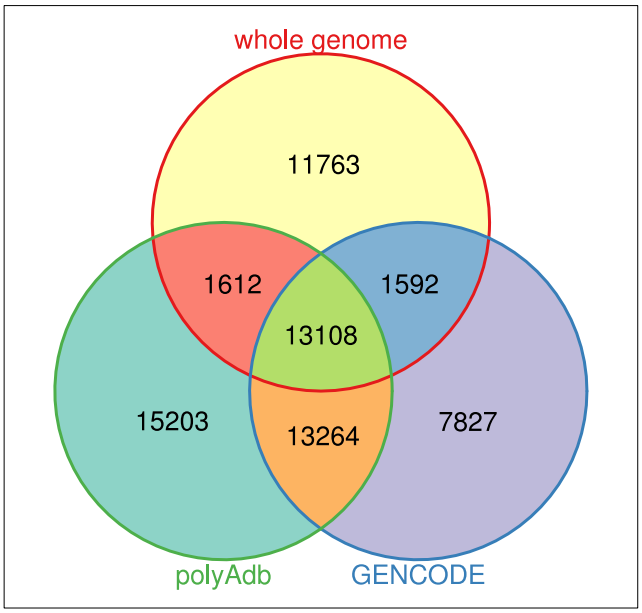

Figure 1: Overlap between identified poly(A) sites and two annotations

Polyadenylation in the nucleus and in the cytoplasm

We merged the poly(A) sites from 5 cell lines in the nucleus and cytoplasm separately for the poly(A)+ and poly(A)- fractions and compared the distribution of the polyadenylation sites in the genome (Figure 3). The difference between poly(A)+ and poly(A)- fractions is that a poly(A) filtering step in the sample preparation protocol assigns RNA with a poly(A) tail on average longer than 30nt to the poly(A)+ fraction, and the remainder goes to the poly(A)- fraction.

Poly(A)- vs poly(A)+ in the nucleus and cytoplasm

By comparing the whole cell data with the nucleic and cytoplasmic compartments in Figure 3, it is clear that the distribution of poly(A) reads in the genome is similar for the poly(A)+ fraction.

fractions From Figure , it can be seen that there are found proportionally a larger fraction of poly(A)- reads in the nucleus than in the cytoplasm. The polyadenylation marks in the cytoplasmic poly(A)- fraction can come from mRNA that is being degraded (as the poly(A) tail is gradually decreased in length before degradation), which will be discussed below.

We wanted to compare poly(A) sites in the nucleus and cytoplasm for different genomic regions. While poly(A) sites in the cytoplasm are expected to derive from from the 3' ends of stable, multi-copy mRNA, the poly(A) sites from the nucleus are expected to stem from mRNA diffusing toward the cytoplasm, mRNA undergoing processing, and possibly some RNA undergoing degradation [2, 3].

As can be seen, the nuclear regions capture polyadenylation signals both in the poly(A)+ and poly(A)- fractions in the intronic regions. Intronic polyadenylation has been identified for many genes [5], which could explain that markers

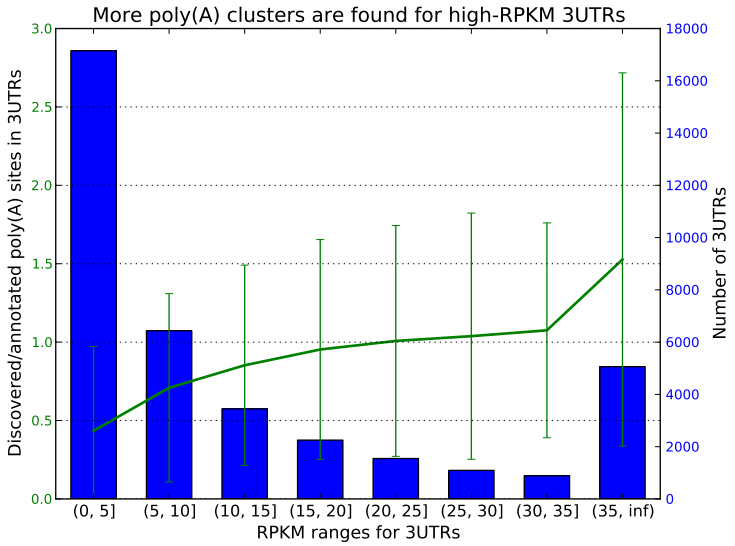

Figure 2: RPKM-poly(A) discovery link

of polyadenylation is found in the poly(A)+ fraction. However, it was unexpected to find polyadenylation markers in the poly(A)- fraction. Their presence in this fractions means that they originate from RNA that had a poly(A) tail of 30nt or less. This could be the evidence of a different type of adenylated RNA in the nucleus, such adenylation as a marker for degradation [3]

It can be seen that for the poly(A)+ fraction the whole-cell samples are most similar to the cytoplasmic ones. However, while the over-all numbers are lower for the nucleus compared to the cytoplasm and whole cell, the relative distribution of poly(A) reads in the different genomic regions is the same. This could indicate that previous studies of polyadenylation that have used whole cell extracts and poly(A)+ fractions have given conclusions that are representative for both the cytoplasm and the nucleus.

In the poly(A)- fraction it can be seen that it is the nucleus that has more marks of polyadenylation than the cytoplasm. In other words, it seems that there is more adenylation with short poly(A) tails in the nucleus compared to the cytoplasm. For some of the adenylation sites in the 3UTR this difference could be explained by polyadenylation caught-in-action, however it does not explain the increase in the other genomic reagions.

To investigate whether the markers of polyadenylation we observe are in the sense or anti-sense direction to anntated genomic features, we looked at only those genomic regions that have features in one sense uniquely, and that do not overlap with any other genomic features. Figure 4 shows this.

Given the recent reports of degradation-related adenylation in humans, we looked for signatures of degradation. Marks of adenlyated assisted degradation can be expected to be less reproducable and less location-specific than 3' mRNA polyadenylation, although in [] the adenylation sites were frequently found at identical or close-by locations both in the nucleus and in the cytoplasm. Any

degrdation-related adenylation site would not be likely to contain the PAS-signal downstream, which 85% of annotated polyadenylation sites in 3UTRs do (Table 1). Thus we searched for adenylation sites that were not at annotated sites and that did not contain one of the PAS downstream. Since degradation-related (A)-tails are shorter than the ones found at the 3' end of mRNAs, these tails are expected to be represented with reads from the 3' pair-end read, thus being originally found with a poly(T) header. Further, as well as genuine poly(A) signals, they are expected to map to the sense strand. Table XXX contains an overview of all polyadenylation sites without PAS. As can be seen .... Is the increase in poly(A)- adenylation in the nucleus to the cytoplasm due to non-PAS adenylation for all regions? Would be great. Yes! You see it. It's not a very string signal, but both the T/A ratio and the non-PAS ratio increases between poly(A)- nucleus and cytoplasm.

Given the possibility of non-PAS adenylation in the human genome, we decided to search for it. Assuming that all reads from 'T' For all the poly(A) clusters we found, we calculated the non-PAS clusters to PAS-clusters (from T-based clusters). Figure X compares cytoplasmic and nucleuc poly(A) minus. As can be seen ...

Nucleus: out of clusters with 2 or more:

Total 3777. With Ts: 2417- 700 (annotated + PAS)

Further: this definitely holds true if we look at the clusters with only 1 coverage! Here there are 23k Ts and 11kAs for NucleusMinus, while Cytoplasm has got 3.5k to 3.7k. Next question is, where do they come from? What genomic region?

obvious follow-up question is then: where do they land? In what types of transcript do they land? coding? noncoding? blablu blibli? Maybe use Andrea's index of gencode 7 for finding out where they land?

Also, the sense/antisense strand would be nice too. Then you have to fix all that. The

DISCUSSION

1. It's possible to use conventional RNA-seq to study polyadenylation, you just need a lot of reads

2. Evidence for geniune poly(A) reads lie in the strandedness of the reads and the A/T origin of the reads. Long tails may have A-endings, but probably few, while short A-tails are expected to have only Ts.

3. Difference in nucleus/cytoplasm for poly(A)+/-, maybe due to degradation-related adenylation?

References

[1] Young Seoub Park, Sang Woo Seo, Seungha Hwang, Hun Su Chu, Jin-Ho Ahn, Tae-Wan Kim, Dong-Myung Kim, and Gyoo Yeol Jung. Design of 5'-untranslated region variants for tunable expression in escherichia coli. *Biochemical and Biophysical Research Communications*, 356(1):136–141, April 2007.

[2] Natalia Shcherbik, Minshi Wang, Yevgeniya R Lapik, Leena Srivastava, and Dimitri G Pestov. Polyadenylation and degradation of incomplete RNA

polymerase i transcripts in mammalian cells. *EMBO Rep*, 11(2):106–111, February 2010.

[3] 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.

[4] 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.

[5] Bin Tian, Zhenhua Pan, and Ju Youn Lee. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Research*, 17(2):156 –165, February 2007.

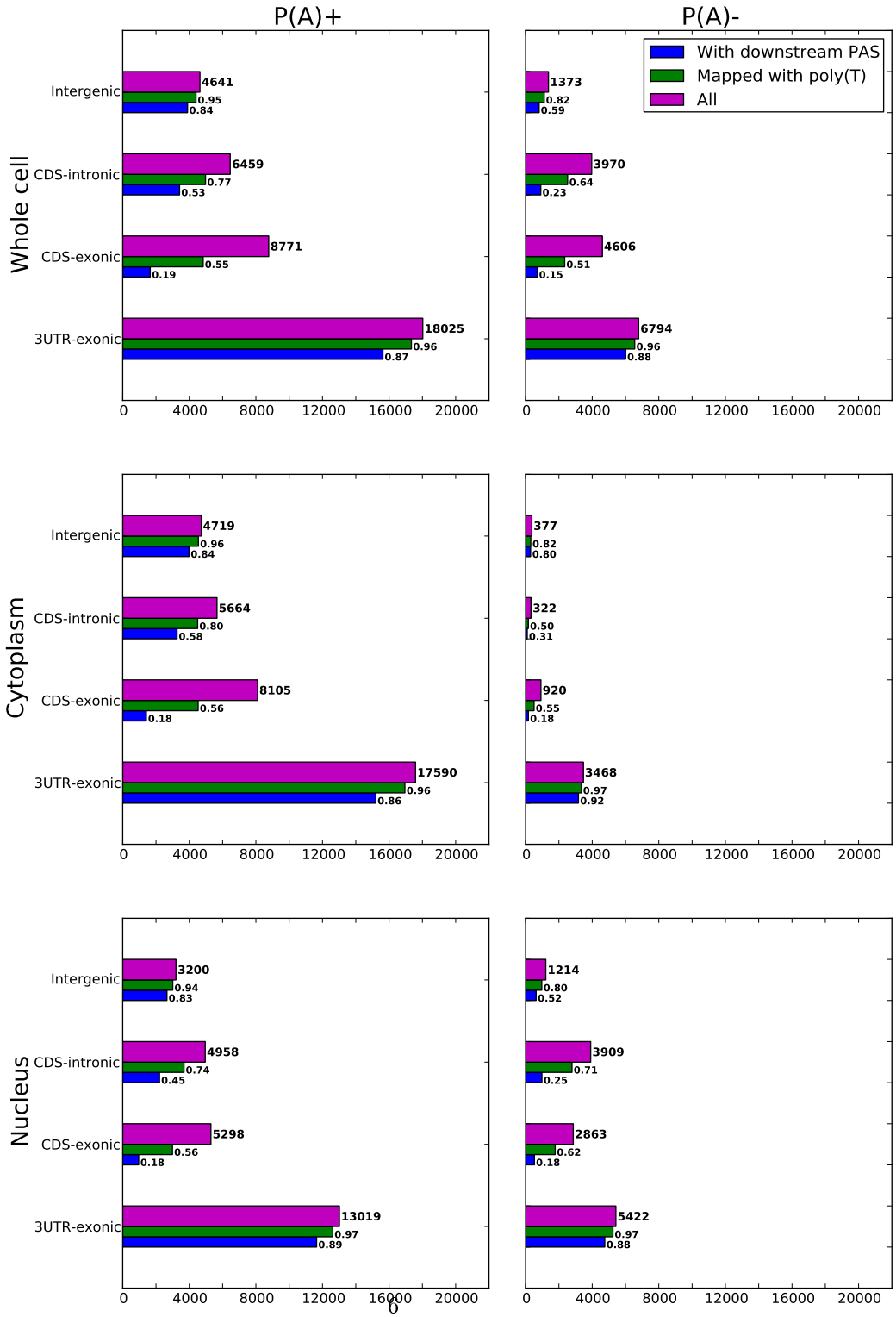

Figure 3: Polyadenylation marks in different cellular compartments, in different genomic regions, for both poly(A)+ and poly(A)- fractions.

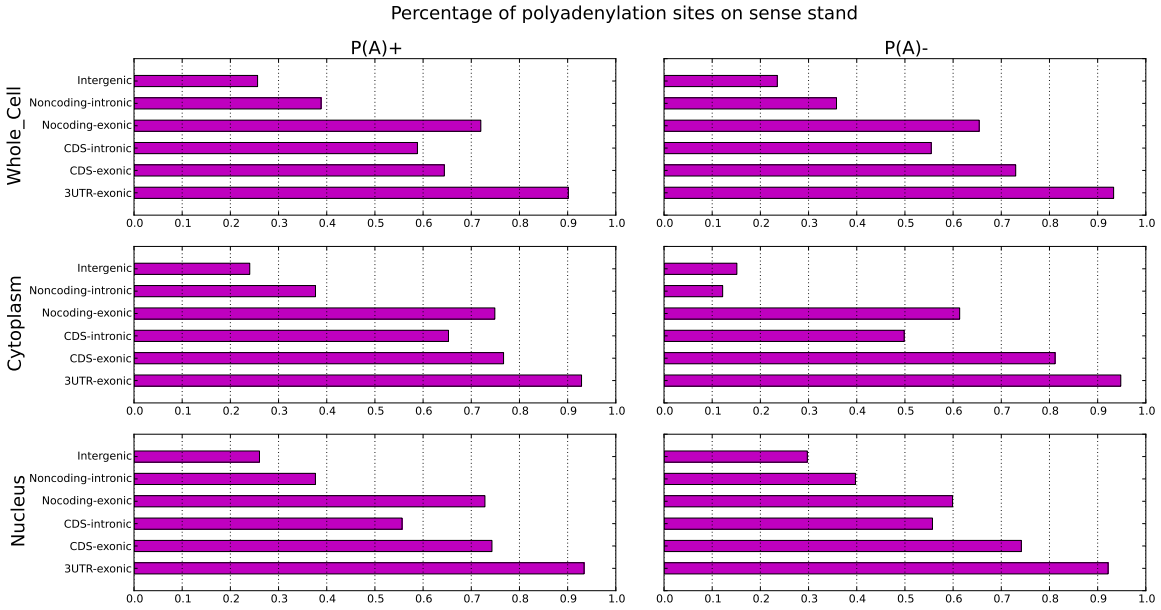

Figure 4: Sense strand mapping percentages in different cellular compartments, in different genomic regions, for both poly(A)+ and poly(A)- fractions