Annotation of Quiescent RNA and Transcription

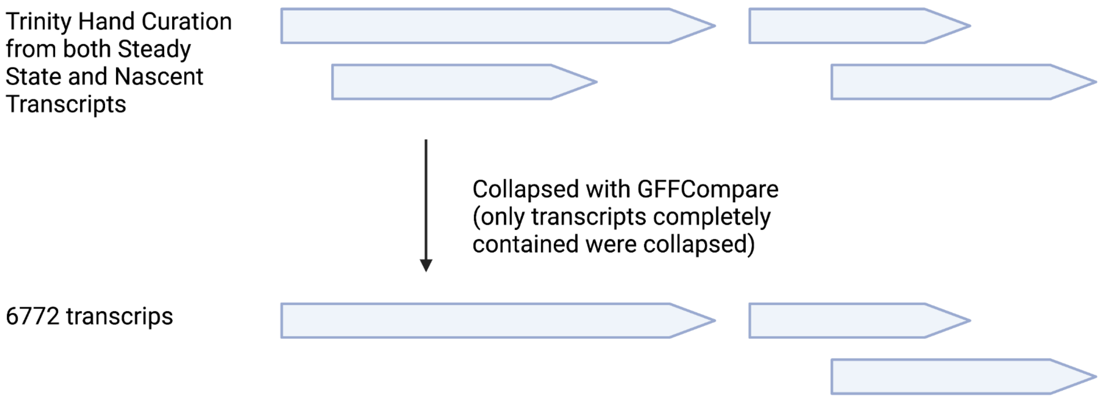
Process in Brief:

Transcripts were generated using Trinity on genome guided mode. Biological replicates were aligned to saccer3 and collapsed. Steady State and Nascent were given to Trinity separately. Trinity constructed transcripts were aligned to saccer3 using gmap and then Steady State and Nascent .gff files were combined using GFF compare.

Transcripts were then hand curated. This mainly involved removing 2 types of transcripts by eye:

1. Expected mRNA transcripts (no change in start or termination site)
2. Incorrect transcripts (in particular Trinity's k-mer approach is particularly bad at repetitive regions)

To make downstream analysis easier, contained transcripts were collapsed with GFFcompare as shown below. For some purposes uncollapsed annotation may be relevant.


Trinity Hand Curation from both Steady State and Nascent Transcripts
Collapsed with GFFCompare (only transcripts completely contained were collapsed)
6772 transcrips

5447 transcripts were then classed as described below.

Transcript Classes

This analysis was done almost entirely with the Bedtools suite. All procedures were executed so that strandedness was forced. Bedtools intersect allows transcripts which overlap between annotation files to be extracted. In this manner, I was able to compare with known annotations.

When compared to a gtf with 14623 annotated transcripts, including:

- mRNA/tRNA/snRNA/tRNA from SGD
- Noncoding/cryptic transcripts:
    - Antisense transcripts - Yassour et al., 2010
    - CUTs, SUTs - Xu et al., 2009
    - CUTs - Vera and Dowell, 2016
    - XUTs - van Dijk et al., 2011
    - SRATs - Venkatesh et al., 2016
    - NUTs - Schulz et al., 2013

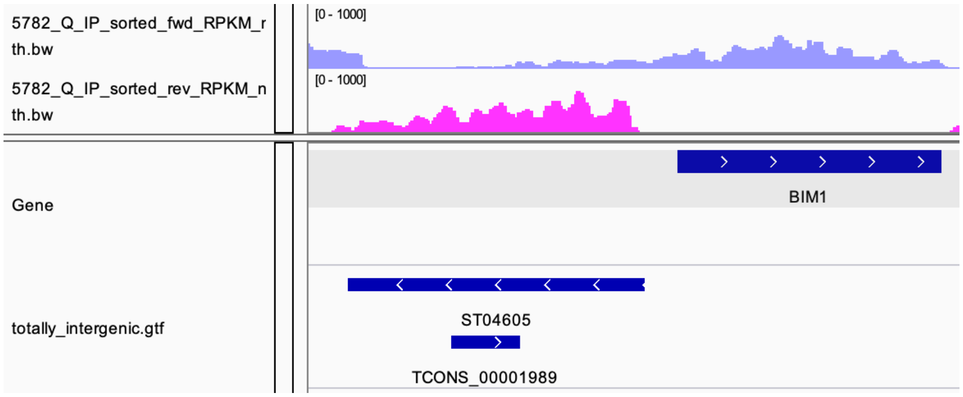
1699 transcripts had no overlap on the same strand with any of the above transcripts. This comes out to ~31% “totally unique”. This is indicative of the necessity of performing this annotation.

mRNAs were extracted and saved separately and then transcripts were compared further in order to be classed.

Intergenic Transcripts

1275 transcripts had no overlap on either strand with annotated mRNAs.

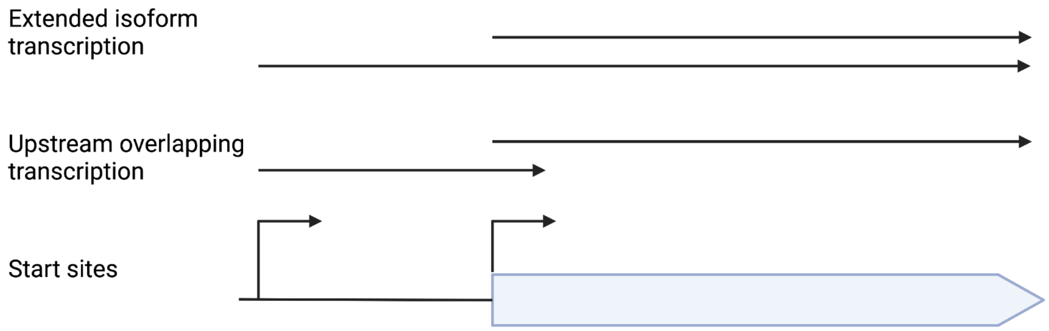
Example (please note bigwigs are RPKM normalized - 1000 = 10 rpkm)



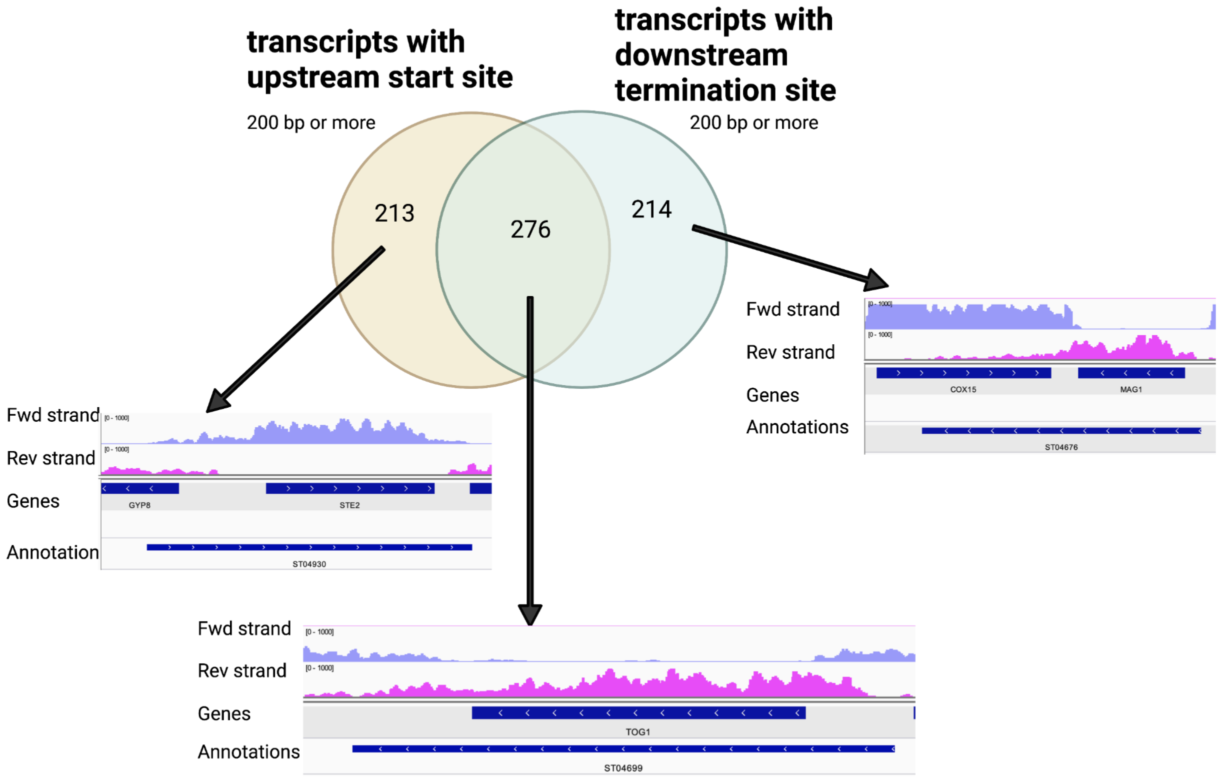
Extended Transcripts:

714 transcripts completely overlapped 100% of an mRNA.
Of those 489 extended 200 bp or more upstream from the annotated start site and 490 extended 200 bp or more downstream from the annotated termination site.

Note: for these transcripts it is not always clear if transcription upstream is part of a unified transcription unit or separate overlapping transcription units. These 2 possible scenarios cannot easily be differentiated by RNAseq alone:


Extended isoform transcription
Upstream overlapping transcription
Start sites

Examples of types of extended transcripts:



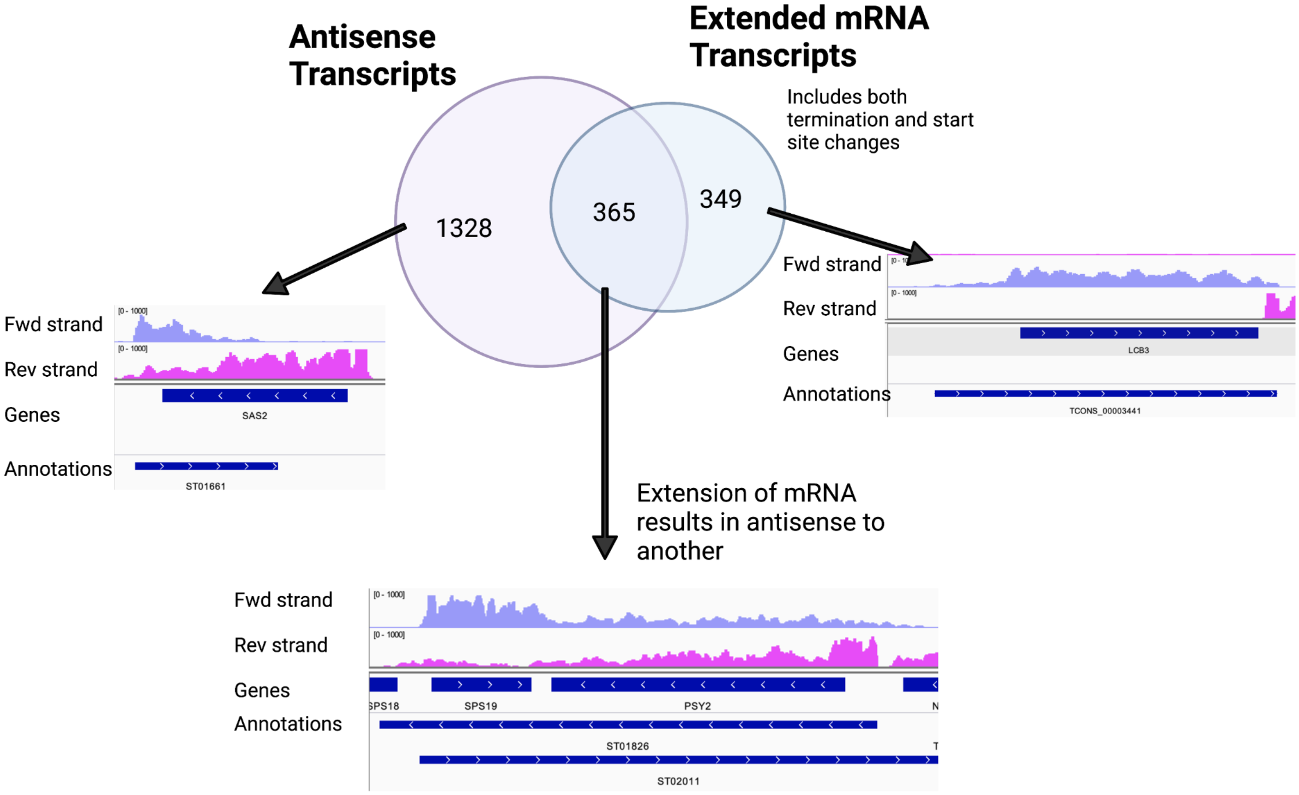
Antisense Transcripts:

To find antisense transcripts -S was used to force antisense strandedness and transcripts were pulled with overlap to mRNA on the opposite strand. Since many antisense transcripts do not extend the full length of the gene I used a percentage cut off.

Antisense to 20% of an mRNA - 2278 antisense transcripts
Antisense to 30% of an mRNA - 1693 antisense transcripts
Antisense to 40% of an mRNA - 1304 antisense transcripts

30% cut off is quite stringent, whereas 20% leads to small fragments that Trinity did not have enough coverage to string together. So while 30% as a cut off likely excludes some real antisense transcripts, 20% includes more noise.

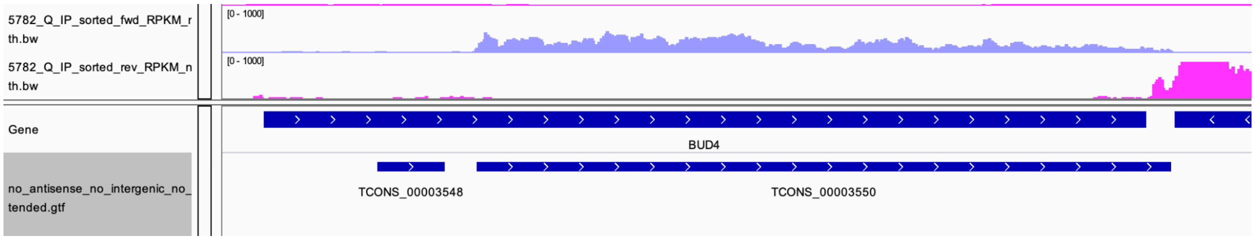
Overlap between Antisense and Extended Transcripts:
Used 30% cut off here



<u>What did not get assigned a class by these metrics?</u>

Using Bedtools subtract, I can pull out all remaining transcripts. This leaves 1607 transcripts which were not classed by this methodology. By eye they include:

- Short antisense transcripts
- Upstream and downstream intergenic transcripts which slightly overlap mRNA
- Small number of sense strand transcripts which initiate intragenically (example below)


5782_Q_IP_sorted_fwd_RPKM_r
th.bw
[0 - 1000]
5782_Q_IP_sorted_rev_RPKM_n
th.bw
[0 - 1000]
Gene
BUD4
no_antisense_no_intergenic_no_
tended.gtf
TCONS_00003548
TCONS_00003550

Expression:

Using DeSeq2, as expected I found that a large percent of these Q transcripts are overexpressed in Q compared to G1. Nascent Transcription and steady state RNA were compared separately.

log2FoldChange>2 and p-value adjusted < .05 were used as cutoffs.

2068 transcripts are up in Q Nascent Transcription as compared to G1 Nascent transcription.