**Goal:**

The Trinity Base assembly need not be Perfect. It is more a matter of if it correctly calls genomic features of interest. All mRNA annotation need not be perfect if it can be filtered out.

However, the better job Trinity does, the less work is needed to make the filtering work well – makes it easier to have logical operators which effectively separate transcription of interest (Antisense RNA, changes in STOP/START sites, intergenic transcription) from known transcription.

What do we want?

This is hard to name. Simply put we want Trinity to be as accurate as possible to the RNA/Transcription that is occurring in the cell. It is hard to know what that *really* means since illuminia short read sequencing doesn’t always make it easy to immediately know what is correct in terms of 3’ and 5’ ends.

The utility of the end assembly is to be able to name the following:

* How many antisense transcripts in Q vs G1 – is what is different? simply level, start site, and/or position?
* How many intergenic transcripts in Q vs G1 – is what is different? simply level, start site, and/or position?
* How many genes have altered transcription start or stop sites in Q?
* How many transcripts do we find that have never been annotated in other conditions – mostly noncoding presumably.
* Additionally, once we have exact coordinates we can use annotation to measure differential expression, or evaluate chromatin structure at particular kinds of loci, etc.
  + Eg. My first annotation attempt allowed me to show that the 3’ NDR at convergent genes that fail to terminate is shallower than convergent genes that terminate appropriately.

**Trinity Annotation Known Issues**

Fragmented Assembly

Issue: Many annotations overlapping at one locus where only one or two make sense.

In some cases, this appears to be a read depth issue. In others is may be the result of our yeast (W303) being genetically quite different from the reference (S288C/BY).

I do wonder if some of this is real, and transcription is really this messy – if there’s any way to know, that would be great.

Chart

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with low confidence

Noisy and confusing annotation in repetitive regions

* Namely TY and (sub?) telomere regions
* Note: Bigwigs were made with minimum mapping quality filter so often it looks like there is nothing here. It is relatively easy to make new bigwigs if you like me to do so without such filters.

A picture containing graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Multiple genes annotated as one

Despite visually appearing to be 2 separate transcription units, sometimes trinity combines genes into one. I think this is a min-glue problem.

**Timeline

Description automatically generated with medium confidence**

**A picture containing graphical user interface

Description automatically generated**

**Filtering after Trinity**

Once we have a suitable trinity assembly, this is how I imagine filtering it. We should talk more about this, but this is the logic that makes sense to me.

1. Junk – take out all the junk and save all it to a gff/gtf file
   1. Anything longer than longest transcript in yeast
   2. Anything with bigger intron than largest known in yeast
   3. Anything with 3 or more exons
2. Filter by minimum expression level
   1. Take out anything below some threshold – will likely need to mess around with this.
3. Split into mRNA vs non-mRNA
   1. Does it overlap more than 80%/90% (?) with a known coding region
      1. If yes mRNA
      2. If no “Bin 2” (presumed noncoding)
4. For mRNA
   1. Where is start site vs. Steinmetz or other annotation – if x upstream or x downstream – pick out
   2. Repeat with 3’ end
   3. How much different than known annotations is a question – I used 200 bp first time but this is likely too strict.
5. For “Bin 2” (noncoding Bin)
   1. Is it antisense to an mRNA
      1. What fraction of an mRNA needs to be covered? I used 30% last time.
      2. Once we have separated AS these can be further classed – More on this later!
   2. Is it not overlapping known mRNAs – intergenic

I have attached 6 pages of bedtools scripts that assess some of these things, though only somewhat systematic. I have others as well but this seemed like definitely enough to start with so you could understand how I went about things. We should definitely talk more about this later but it might be too much detail now. These were rough notes mostly for my own benefit, so feel free to ask any questions.

**Misc Other Questions**

* Is 4tU-seq signal, be it nascent and/or stead-state, more like, say, ChIP-seq signal than RNA-seq signal?
  + no, it is more like RNA-seq
* I was assuming that steady-state 4tu-seq signal is the same as RNA-seq signal, but is that assumption correct? If not, then how are signals from each experiment different and in what ways?
  + Mostly! More handling happens so Input and steady state might be slightly different – especially Steady State of unlabeled cells, but we assume they are.