Class 18: RNAseq

Reading assignment

- Nagalakshmi et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. 2008 320:1344.

Classroom activity (limit 45 minutes)

1. This paper was published about 10 years after the completion of the yeast genome sequence. What were the continuing challenges of genome annotation does this paper identify? The paper describes the authors invention of the RNAseq procedure that has become so popular. Describe the procedure used by the authors.

   The issue with annotating was to accurately identify the boundaries of genes and transcription units. The question remained as to how many genes the genome included and what proteins those genes were predicted to encode. RNAseq held the promise of defining both much more accurately. The procedure involved isolating poly(A) RNA (by definition, mostly mRNA), generating cDNAs by priming with oligo(dT) or random hexamers (6 nt primers), fragmenting the ds cDNA into small (undefined in print) fragments, then sequencing about 35 nt from each end by Illumina sequencing.

2. Given the numbers quoted for length and number of sequence reads, how many nucleotides of yeast DNA sequence were identified in this experiment (with the hexamer or oligo(dT)-primed cDNAs)? The yeast genome is ~12 megabases, so if the RNAseq results were random across the genome, how many times would each nucleotide be represented on average? Is that overrepresentation important? Why?

   16 million (hexamer) and 14 million (oligo(dT)) sequences of 35 nt each were determined. That corresponds to 560 and 490 million nt. Given a 12 million base genome, that would correspond to averages of 47 and 40-fold overrepresentation. Since only 75% of the genome was represented, the real overrepresentation would be 63 and 53-fold. The overrepresentation was important to provide as redundant a set of data as possible to make it possible to draw the conclusions described in the rest of the paper. The differences in data at different genomic locations would not be due to randomness in the data given its redundancy.

3. How did the authors assess their RNAseq results to determine their validity? Why was it so important for them to do that? How well do these tests demonstrate validity of the data?

   The genome they sequenced included a 3.5 kb deletion of the LEU2 *gene and none of the sequences matched to that region, which is an indication of the specificity of the sequencing results. Second, they performed two technical replicates (retesting identical cDNA samples) of each biological replicate (sequencing duplicate cDNA samples) were tested and all the replicates gave extremely similar results (Pearson correlation coefficients of 0.93 to 0.95 for biological replicates and 0.99 for technical replicates). The second test shows that there is little variability in the process so even single RNAseq results should be believed. The first is a weak test because even if anomalous sequencing results occurred (like reverse transcriptase jumping from one RNA template to another, juxtaposing sequences not found in the genome) there would be no sequences of the* LEU2 *region.*

4. What aspects of the structure of the transcriptome (or of its constituent mRNAs) did the RNAseq data demonstrate? Provide an overview of that data.

   The data showed that about 75% of the genome is transcribed even though the more of the sequences near the 3' ends of mRNAs were represented (this may affect later the issue of determining the relative amounts of mRNAs below). They were able to map the 5' and 3' ends of the mRNAs finding the average 5' UTR to be 50 bp (varying from 0 to 990 bp). 241 genes had an ATG within 10 bp of the 5' end and they questioned if these could be used as

initiation codons. 3' UTRs averaging 104 bp (varying from 0 to 1461 bp). 275 pairs of transcripts overlapped at their 3' ends (12% of the total) suggesting a possible novel form of regulation in yeast involving pairing between 3' ends of transcripts.

5. The authors describe two features having to do with the initiation codon, ATG, in the yeast genome. What is the difference between stating the number of genes with additional ATGs that extend the previously predicted ORF and the existence of upstream uORFs? Given what you know about the process of identifying initiation start codons in eukaryotes, in which case would the additional ATG have the effect of reducing expression of the downstream gene?

   In the first case, the ORF that defines the gene had been recognized as starting at the first of the two ATGs and the RNAseq data now shows that the ORF could (but might not) begin at an upstream ATG. The result is that the protein would be somewhat longer, but no other change in expression. In the second case, an ATG is upstream of the recognized start site, but it is at the beginning of an ORF upstream of the gene (a "uORF") and so should specify a short peptide. The issue for translation of the gene is that in a eukaryote translation starts at the first ATG after the 5' end and so to the extend that the uORF is translated the downstream ATG cannot be recognized as a start site. There are complications to this view that we could discuss if we wish.

6. The authors describe whether RNAseq validated annotated introns in the yeast transcriptome. What was the signal in the RNAseq data for the inability to validate these introns? Were they able to confirm all previously identified introns? They also searched for evidence of transcription in intergenic regions. What was the evidence for this type of transcription? How did they test this observation? Was the observation confirmed?

   They identified the site of introns that had been spliced as a discontinuous sequence (not present in the genome) flanked by the conserved GT/AG sequences immediately upstream and downstream of the intron. Surprisingly, they found no evidence of such a linkage in many cases. Most of these were from non-transcribed genes but of 30 others only four could be shown to not be spliced. They looked for transcription of intergenic regions and found 487 cases, 204 of which were not previously seen with microarrays. They did qRT-PCR on a sample of these and confirmed most. This shows a significant amount of intergenic transcription does occur in yeast. Both pieces of data show that RNAseq is more sensitive than previously used methods and therefore is preferable.

7. In a final comment, the authors describe the use of RNAseq to measure RNA expression levels. How well did RNAseq correspond to other whole genome methods of detecting RNA? What advantages does RNAseq provide compared with these other methods?

   At the end of the paper they suggest this possibility and showed that the results of RNAseq were strongly correlated to results of qRT-PCR for genes transcribed at high and moderate levels. The correlation was less for those transcribed at low levels. They listed several advantages of RNAseq: it has a huge 8000-fold dynamic range compared with about 60-fold for microarrays; it allows testing of all unique sequences, rather than only those specifically tested in microarrays; and RNAseq allows accurate assignment of all boundaries in genes including 5' and 3' ends and intron boundaries, whereas microarrays only allow an approximate assignment.