non-coding RNA analysis

Non-coding RNA analysis

Non-coding RNA analysis can follow one of three paths:

1. Using a reference (usually genome) for analysis of known ncRNA or de novo discovery
    a. Step 1: Map the RNA data to a reference; you must chose whether you will or will not allow for editing since this will determine the type of mapper you must use
    b. Step 2: Interpret the mapping results
        i. For known ncRNA, a tool such as Tophat will fold together the mapping results with the annotations provided and give you, e.g., counts of ncRNA.
        ii. For discovery of novel ncRNA (like miRNAs) usually this entails extracting the genomic region covered by some non-coding RNA and using auxiliary tools to analyze that region. For example, if you find an abundance of 18-30 nt RNA in a tight region, you might extract +/- 100 bp from that region and pass that to an RNA folding program like `randfold` to see if they form a hairpin. If you've separated RNA fractions by size, you could look for separate evidence of mature and precusor RNA.
2. Not using a reference
    a. Abundance based methods: boot-strap pseudo-assemble ncRNA's (or if they're very short form a consensus sequence) based on the raw abundances of reads. Should work well if read length is larger than your expected ncRNA size.
    b. Assembly based methods: tweak and tune your favorite assembler to try to assemble short contigs. Tricky part is picking k-mer size.

As a local resource, the Sullivan Lab at UT has become quite expert at ncRNA analysis, particularly of miRNA's in viral systems.

Interpreting the results

Homology, homology, homology. Some suggestions:
a) ncRNA populations tend to be highly skewed; consider methods that "subtract out" reads by mapping; for example if you have contaminating rRNA fragments, map to a close rRNA reference and remove those reads, proceed to some subpopulation like snoRNA's, remove those, etc.
b) try to use only the data itself first based on abundance, then look for homology to the abundant sequences; if they are fragments which then assemble properly, your homology search before and after assembly should agree.

Gotchas

1. Remember that short ncRNA will almost certainly have adaptor sequence in it which you'll have to trim out. This can be to your advantage since it gives you the precise endpoint of the RNA from single-end sequencing data.
2. Watch out for degradation products, especially from tRNA and rRNA which may be much more abundant than the interesting ncRNA's you're looking for.