<https://davetang.org/muse/2012/05/25/how-to-deal-with-multi-mapping-reads/>

**A survey of best practices for RNA-seq data (2016)**

Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples, but it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads.

What do they recommend:

1. Kallisto (2016)
2. RSEM (2011)
3. HTSeq (2014)
4. eXpress (2012)
5. ShortStack (but more so for sRNA mapping and characterization)

RPKM, FPKM, and TPM normalize away the most important factor for comparing samples, which is sequencing depth.  These approaches rely on normalizing methods that are based on total or effective counts, and tend to perform poorly when samples have heterogeneous transcript distributions, that is, when highly and differentially expressed features can skew the count distribution,

K, so what do we have at this point to try and explore:

1. MuMRescueLite (2009) (90cites)
2. Rescue Strategy for multi-mapping reads (2009) (90cites)
3. RSEM (2011)
4. HTSeq (2014)
5. eXpress (2012)
6. ShortStack/Butter (2016)
7. Kallisto (2016)

**Improved Placement of Multi-mapping Small RNAs**

For example, the central miRNA database, miRBase, is thought to contain substantial numbers of incorrect annotations (Kozomara and Griffiths- Jones 2014; Taylor et al. 2014).

A major issue is the prevalence of multi-mapping (MMAP) reads in sRNA-seq data. MMAP reads occur when there are multiple best-scoring alignments to the reference genome.

MMAP reads are quite rare in modern polyA+ mRNA-seq data due to their longer read-lengths, and due to the fact that polyA+ mRNAs generally are transcribed from single-copy sequences.

MMAP reads are much more frequent in sRNAseq data due both to the short lengths of the reads, and their tendency to originate from higher-copy number regions of the genome. Note citations from etc,

MMAP sRNA-seq reads are often dealt with simplistically, either by randomly selecting one the possible alignment positions, or by ignoring them entirely. Here we see the conflicting issue of precision(choosing an exact alignment) vs sensitivity(taking into account information of all alignments.

Here, we describe the implementation and testing of sRNA-seq alignment performance using a local-weighting method to better place MMAP small RNA reads.

**Methodology and Reasoning**

1. We assume that each read in a sRNA-seq experiment represents a single small RNA molecule, which therefore must have had a single genomic origin. It’s quite a loose and duh assumption, basically that a single read must have necessarily come from one place on the genome. So if we have 100 reads of identical sequence, and MMAP=2, it is entirely possible that some came from 1, the rest from the other.
2. Our initial goal of siRNA seq would be to identify identify the site of transcriptional origin of the small RNAs, not to list their possible targets. Target identification comes later.

Methodology?

ShortStack uses bowtie (Langmead et al. 2009) to identify all possible best-matched alignments for each read, subject to a default, user-adjustable limit of 50 alignments per read.

ShortStack will then calculate a probability for each alignment according to one of three alternative methods.

Following the example, we have a read with MMAP=2.

Each of the two possible loci in this example have different “neighborhoods” of adjacent alignments of other reads

1. Randomly assign. Each position results in 50-50
2. Unique-weighting. Given all the mapped regions, calculate total unique mappers within multiple regions. Then use that number as denominator. Each MMAP position is then weighed by dividing counts of unique-mappers/denominator.
3. Fractional-weighting. Calculate the ‘score’ for each read in the vicinity, inclusive of the read in question. This takes into account multiple mappers in the vicinity. The higher the number of unique mappers obviously the higher total region score. The higher the MMAP value, the lower it contributes to score. Then each region’s score is expressed as a fraction of (all region) score.

Genome is divided into 50-nt bins, and region is defined as a 5xbin ie, 250nt window centered around the alignment.

The calculated probabilities are then used as weightings in a random number selection to designate the primary alignment for the read (Figure 2F), and all of the other possible alignments are marked as secondary alignments.

**Important notes**

Also should note, that in the given example, there is still a 2.5% probability that the less-likely position would be marked as the primary alignment.

Also the analysis is at a per-read basis, not a per-sequence basis. Therefore, 100 reads of 80%-20%, MMAP=2, 80 of those reads will have primary alignments at pos1, 20 of those reads at pos2.

**Experimental Testing**

***Arabidopsis miRNA***

Mature miRNAs are frequently encoded by multiple paralogous loci, which causes sRNA-seqreads corresponding to the mature miRNA to be MMAPed.

However, the primary transcripts derived from each of the paralogs typically are much longer, and have distinct sequences, such that qRT-PCR approaches can distinguish them, allowing use to assess true expression levels of each transcript.

Three miRNA families were studied, each with varying number of paralogs and with each paralog having varying expression levels.

Seems like ShortStack’s U method produced mature miRNA alignments that were mostly similar to the results from qRT-PCR of the primary transcripts, whereas the other methods resulted in alignments that suggested relatively equal levels of expression among paralogs.

***Arabidopsis ds-siRNA***

Similar results were shown , U was shown to have higher precision.

Flaws of bowtie? Preferentially selects plus over minus strand.

A particularly important goal of ours is to reduce the likelihood of false

discoveries in small RNA gene annotation. Short-

Stack-U gives fewer false positives for MMAP reads than all other

methods for sRNA-seq alignment (Figure 4, Figure 5, and Figure 7)

Now lets think about this? Why does U give more accurate results? By only taking into account unique mappers? Having a higher number of unique mappers in a region represents greater confidence/ certainty that that region should be given greater weightage? That that region truly has a higher level of expression?

Is it because these MMAP regions only constitute a part of the region/ gene, therefore, unique mappers gives us greater certainty than multiple mappers.