Bioinformatics

Difference between de novo transcriptome assembly methods

Asked 5 years ago Modified 5 years ago Viewed 1k times

5

I have been looking around (including read the original papers) to understand what is essentially the difference between StringTie in non-reference based mode (de novo) and Trinity de novo assembly. I understand that in the genome-guided nature of StringTie, we are supported by the genome information (i.e. taking into account the mapped reads). In addition to that, we also have an option to either run using the gene annotation file or not.

Simply put, am I correct if I say that Trinity is purely looking at the overlapping reads, then assemble them, whereas StringTie (non-reference based) is looking at the mapped reads that are either close or overlapping to each other? Is there any other thing I'm missing here (I understand it's not that simple, but I'm trying to make it as simple/intuitive as possible)?

Also if that's the case, would it be safe to say that generally non-reference based StringTie is preferred over Trinity and reference based StringTie (as gene annotation can lead to some "bias", hence preventing discovery of novel TSS, exon length or even make unnecessary duplicates with small differences)?

rna-seq sequence-annotation assembly transcriptome

Share Improve this question Follow

asked Jan 3, 2018 at 15:25
kaka01
111 1 6

2 Answers

Sorted by:
Highest score (default)

3

Your intuition is correct. stringTie is just looking at clumps of alignments and how they might relate to each other (either due to spliced alignments or proximity). Trinity is doing the more computationally difficult task of finding parts of reads that overlap with each other and trying to link them together into longer sequences.

Whether it makes sense to use something like stringTie or not depends mostly on the quality of the reference sequence. If you're working on an organism with a high-quality reference sequence (pretty much anything with a lot of papers using it) then there's no reason to bother with Trinity. If your reference genome additionally has a high-quality annotation (human, mouse, etc.) then you should use it, as anything "novel" that the program is likely to output otherwise is probably just an artefact. If you either don't have an annotation or it's known to be highly problematic then it makes sense to omit it when using stringTie.

Without seeing evidence to the contrary, I don't buy the argument that using a good annotation is actually biasing the output in a problematic way. For well-studied organisms, anything reported as "novel" these days likely doesn't exist, or is otherwise so fleeting that it's biologically irrelevant. The arguments of bias sound like something that was come up with once someone didn't get as many "novel" results as they were hoping.

Share Improve this answer Follow

answered Jan 3, 2018 at 15:53
Devon Ryan
19.3k 2 27 58

Thanks for your answer. About the point we will less likely find novel things in organism with high-quality annotation, I was wondering if that's always the case as there are many recent publications highlighting novel elements, even in human. I guess you are saying that generally it's sufficient, but in the case of some specific condition or localisation, finding novel elements is still relevant? – kaka01 Jan 4, 2018 at 12:34

1 Finding a few of them, sure, but unless the system is being severely disturbed one would only expect a handful of real novel findings (those reporting more are likely primarily false-positives at this point). – Devon Ryan Jan 4, 2018 at 12:42

3

It's really a misnomer to call StringTie's non-reference based mode 'de-novo.' It's still using the reference genome sequence to guide the transcript assembly, it's just not using the reference annotation. Trinity is truly de-novo in that it it assembles the transcript from the overlap of the reads without mapping them to a reference genome sequence.

If you look at the precision numbers from the StringTie paper, you'll see that it maxes out around 80% under ideal circumstances with simulated data. This means that if you're working with a well annotated genome where most transcripts have already been discovered, the vast majority of the 'novel' transcripts you find will likely turn out to be artifacts.

Share Improve this answer Follow

answered Jan 3, 2018 at 16:06
heathobrien
1,806 6 16

Thanks for the answer. I agree for the first point. For the second point, I partially agree as I think we can always find some novel transcripts in the case of uncovered condition/localisation (as my comment to Devon's answer). I hope "vast majority" here means that there could still be a few out of thousands that are indeed biologically relevant. ;) – kaka01 Jan 4, 2018 at 12:39

1 Happy to help. It's certainly possible that some novel transcripts will be real, especially if your experimental setup is novel, but validation is likely to be a major bottleneck. In my opinion, short reads are really only suitable for quantification of known transcripts. Transcript discovery is much more

effective with long read sequencing technologies like PACBIO Iso-Seq or Oxford Nanopre. – heathobrien Jan 4, 2018 at 14:22