A decade of de novo transcriptome assembly: Are we there yet?

Martin Hölzer[1]

[1]Robert Koch Institute

September 11, 2020

Abstract

A decade ago, de novo transcriptome assembly evolved as a versatile and powerful approach to make evolutionary assumptions, analyze gene expression, and annotate novel transcripts, in particular, for non-model organisms lacking an appropriate reference genome. Various tools have been developed to generate a transcriptome assembly, and even more computational methods depend on the results of these tools for further downstream analyses. In this issue of Molecular Ecology Resources, Freedman et al. (2020) present a comprehensive analysis of errors in de novo transcriptome assemblies across public data sets and different assembly methods. They focus on two implicit assumptions that are often violated: First, the assembly presents an unbiased view of the transcriptome. Second, the expression estimates derived from the assembly are reasonable, albeit noisy, approximations of the relative frequency of expressed transcripts. They show that appropriate filtering can reduce this bias but can also lead to the loss of a reasonable number of highly expressed transcripts. Thus, to partly alleviate the noise in expression estimates, they propose a new normalization method called length-rescaled CPM. Remarkably, the authors found considerable distortions at the nucleotide level, which leads to an underestimation of diversity in transcriptome assemblies. The study by Freedman et al. clearly shows that we have not yet reached "high-quality" in the field of transcriptome assembly. Above all, it helps researchers be aware of these problems and filter and interpret their transcriptome assembly data appropriately and with caution.

A decade of *de novo* transcriptome assembly: Are we there yet?

Martin Hölzer[1-3]

[1] MF1 Bioinformatics, Robert Koch Institute, 13353 Berlin, Germany

[2] RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University, 07743 Jena, Germany

[3] European Virus Bioinformatics Center, Friedrich Schiller University, 07743 Jena, Germany

A decade ago, *de novo* transcriptome assembly evolved as a versatile and powerful approach to make evolutionary assumptions, analyze gene expression, and annotate novel transcripts, in particular, for non-model organisms lacking an appropriate reference genome. Various tools have been developed to generate a transcriptome assembly, and even more computational methods depend on the results of these tools for further downstream analyses. In this issue of Molecular Ecology Resources, Freedman *et al.* (2020) present a comprehensive analysis of errors in *de novo* transcriptome assemblies across public data sets and different assembly methods. They focus on two implicit assumptions that are often violated: First, the assembly presents an unbiased view of the transcriptome. Second, the expression estimates derived from the assembly are reasonable, albeit noisy, approximations of the relative frequency of expressed transcripts. They show that appropriate filtering can reduce this bias but can also lead to the loss of a reasonable number of highly expressed transcripts. Thus, to partly alleviate the noise in expression estimates, they propose a new normalization method called length-rescaled CPM. Remarkably, the authors found considerable distortions at the nucleotide level, which leads

to an underestimation of diversity in transcriptome assemblies. The study by Freedman *et al.* clearly shows that we have not yet reached "high-quality" in the field of transcriptome assembly. Above all, it helps researchers be aware of these problems and filter and interpret their transcriptome assembly data appropriately and with caution.

In software development, it usually doesn't take long for an approach to work fundamentally, or at least to look like it will work. However, there is always a lot of work still to be done "behind the scenes" to catch edge cases, deal with errors, and to find and fix all the little bugs. Accordingly, a rule of thumb, derived from the Pareto principle named after the economist and philosopher Vilfredo Federico Damaso Pareto, states that the last 20% of a software project usually takes 80% of the time.

This rule of thumb can also be applied to many areas of bioinformatics. The massive parallel sequencing of DNA has led to the creation of new genomes, which are being assembled, annotated, and analyzed more rapidly than ever before. However, we are still struggling to get it right (the last 20%, so to speak), as Steven Salzberg noted in a recent report on pervasive assembly and annotation errors (Salzberg, 2019). The same, if not worse, applies to the analysis of high-throughput transcriptome sequencing data (RNA-Seq), where (*de novo*) assembly is a prominent first analysis step. While the assembly of transcriptomes has become an everyday bioinformatics task, dealing with all the potential errors and small caveats is still a challenge and error-prone, even a decade after the emergence of the first tools (Birol *et al.*, 2009; Grabherr *et al.*, 2011; Schulz *et al.*, 2012).

In their recent study, Freedman *et al* . extensively analyzed errors, bias, and noise in *de novo* transcriptome assemblies. In its most common application, RNA-Seq short reads are aligned to a reference genome (map-to-reference, as Freedman *et al* . refer to it) to functionally annotate genomic features (such as genes) and estimate their expression levels. In another application, RNA-Seq-derived reads can be (*de novo*) assembled first to reconstruct the transcriptome and then use it as a proxy for annotation and expression evaluation (map-to-transcriptome).

According to Freedman *et al* ., *de novo* transcriptome assembly is based on two implicit assumptions. First, the assembled sequences represent an unbiased view of the underlying expressed transcriptome, and second, the expression estimates of the assembly are good, if noisy, approximations of the relative frequency of expressed transcripts (Freedman, Clamp and Sackton, 2020). It is evident that these two assumptions have important implications for further downstream analysis steps and directly affect gene expression estimates, variant invocation, and evolutionary analyses based on a *de novo* transcriptome assembly. In their work, Freedman *et al* . show that these assumptions are frequently violated across different public mice RNA-Seq data sets and assembly algorithms, thus directly impacting downstream analyses performed on *de novo* transcriptome assemblies. In particular, they focused on expression estimation bias and differences in nucleotide variant calls while also comparing *de novo* results against a map-to-reference approach.

Firstly, Freedman *et al* . dispel the illusion that *de novo*transcriptome assemblies are mainly composed of full-length transcripts, which is typically not the case for short reads. The authors continue to carry out that the functional composition of a transcriptome assembly is biased towards intronic, UTR, and intergenic sequences, although most studies focus on protein-coding genes. As an important finding, they describe frequent genotyping error rates ranging from 30% to 83% that, in particular, negatively bias heterozygosity estimates (Fig. 1). Their results also show that single contigs are poor expression estimators. Although commonly done in the current gene expression literature, the use of single contigs as proxies for gene-level expression appears to be problematic according to their study. Based on their results, it might be interesting to investigate whether cluster- or graph-based expression estimates can overcome such limitations.

Alongside these interesting, but also alarming findings, Freedman*et al* . suggest ways to deal with individual errors and minimize them. Among other ideas, they propose a new formula for normalizing gene expression, the length-rescaled CPM (counts per million). It is best practice in transcriptomics to consider measures like sequencing depth and feature lengths when estimating and comparing expression values derived from RNA-Seq counts. However, correctly determining a feature's length from a *de novo* transcriptome assembly

alone can be difficult because gene lengths are not adequately represented on the fragmented gene models that are typically derived from *de novo*transcriptome assemblies. To account for such biases, the authors investigated whether rescaling of CPM using length metrics based on information from both reference transcripts (observed length) and*de novo* assembled contigs (effective length) improves expression estimates. By combining effective and observed length, they adjust the CPM values to better represent the actual transcriptome expression. They show that, to some extent, the expression bias at gene level can be corrected by this formula. However, the observed length estimation is difficult for non-model organisms lacking a good reference genome or transcriptome and annotation.

So, are we there yet? With the transcriptome assembly methods for short RNA-Seq reads developed over the last decade, we are quite close to the first 80%. However, as Freedman *et al* . impressively demonstrate, the last 20% still pose a challenge. Multiple tools and parameter settings are often used and merged to generate a comprehensive *de novo* transcriptome assembly, but further bias and redundancy are introduced that researchers need to deal with (Hölzer and Marz, 2019). Nevertheless, modern multi-tool ensemble approaches for *de novo*transcriptome assembly achieve promising results (Voshall *et al.*, 2020). However, the implicit assumptions and their violation, as discussed extensively by Freedman *et al* ., urgently require control mechanisms and corresponding normalization and filter steps, especially with such combined approaches.

Finally, Freedman *et al* . give a brief outlook on the application of long reads derived from single-molecule real-time sequencing (SMRT), as provided e.g. by PacBio or Oxford Nanopore Technologies (ONT), to generate a provisional genome assembly in the absence of a suitable reference genome. Such a draft can then be used for map-to-reference transcriptome analyses. However, other problems may arise, and, as Freedman *et al* . describe, genome assembly is not necessarily a panacea for all issues related to expression analysis.

With a view of today's technology, one could even argue that the transcriptome assembly of short reads will become obsolete in the coming years. SMRT is already capable of generating long reads that can potentially span full-length transcripts - no assembly required!? In addition, ONT allows for the direct sequencing of native RNA molecules (dRNA-Seq) without any fragmentation steps and cDNA conversion. Recently, the application of ONT dRNA-Seq for the detection of differential expression of human cell populations impressively showed the potential of the technology to overcome many limitations of short and long cDNA sequencing methods (Gleeson *et al.*, 2020). However, even with the complete avoidance of biases introduced by *de novo* transcriptome assembly of short reads, not all problems are immediately solved by switching to another technology. Instead, other noise classes occur, such as a higher sequencing error rate for dRNA-Seq, which researchers need to know and which must be taken into account by novel tools. Thus, hybrid approaches combining the strengths of both short and long reads will become more important, in particular in the context of *de novo* assembly and transcriptome analyses. In any case, one thing will certainly not let us go: the careful handling of transcriptome data and their interpretation with regard to error, noise, and bias.

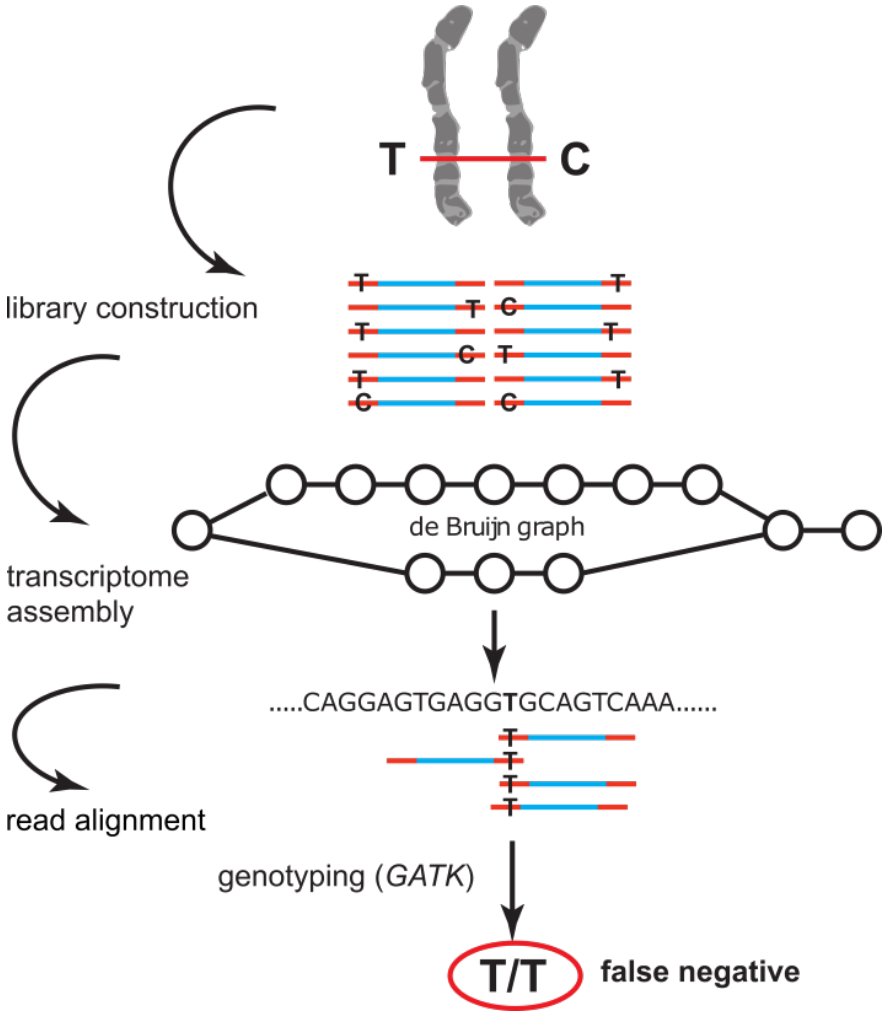

Figure 1: An important result of Freedman *et al* . are genotyping errors based on *de novo* transcriptome assemblies, especially those that, as shown here schematically for a diploid mouse chromosome, tend to underestimate heterozygosity.

References

Birol, I. *et al.* (2009)*De novo* transcriptome assembly with ABySS, *Bioinformatics*, 25(21), pp. 2872–2877.

Freedman, A. H., Clamp, M. and Sackton, T. B. (2020) Error, noise and bias in *de novo*transcriptome assemblies, *Molecular ecology resources*. doi: 10.1111/1755-0998.13156.

Gleeson, J. *et al.*(2020) Nanopore direct RNA sequencing detects differential expression between human cell populations. doi: 10.1101/2020.08.02.232785.

Grabherr, M. G. *et al.*(2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nature biotechnology*, 29(7), pp. 644–652.

Hölzer, M. and Marz, M. (2019)*De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers, *GigaScience*, 8(5). doi: 10.1093/gigascience/giz039.

Salzberg, S. L. (2019) Next-generation genome annotation: we still struggle to get it right,*Genome biology*, 20(1), p. 92.

Schulz, M. H. *et al.*(2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels, *Bioinformatics* , 28(8), pp. 1086–1092.

Voshall, A. *et al.*(2020) A consensus-based ensemble approach to improve *de novo*transcriptome assembly. doi: 10.1101/2020.06.08.139964.