The proper care and feeding of your transcriptome

Richard Smith-Unna[1], Matthew D MacManes [2],

1 University of Cambridge

2 Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham NH, USA

∗ **E-mail: macmanes@gmail.com, @PeroMHC**

Abstract

Some abstract

Introduction

For biologists interested in understanding the relationship between fitness, genotype, and phenotype, modern sequencing technologies provide for an unprecedented opportunity to gain a deep understanding of genome level processes that together, underlie adaptation. Transcriptome sequencing has been particularly influential, and as a direct result, a diverse toolset for the assembly and analysis of transcriptome exists. Notable amongst the wide array of tools include several for quality visualization (FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) and SolexaQA (1)) read trimming (e.g. Trimmomatic (2) and Cutadapt (3)), read normalization (khmer (4)), assembly (Trinity (5), SOAPdenovoTrans (6)) and assembly verificaton (transrate `https://github.com/Blahah/transrate` and RSEM-eval).

mostly because of the relative ease with which they can be produced. This ease in construction, however, does not appropriately reflect the subtle (and not so subtle) methodological challenges associated with transcriptome reconstruction. Amongst the most challenging include isoform reconstruction, simultaneous assembly of low- and high-coverage transcripts, and [], which together make good transcriptome assembly really difficult.

Methodological abuse is widespread. Particularly flagrant are abuses related to quality control of input data, the lack of understanding the role *kmer* selection may play in accurate reconstruction, and lastly, abuses related to the lack of post-assembly quality evaluation. Here, we aim to define a set of evidence based analyses and methods aimed at improving transcriptome assembly, which in turn has significant effects on all downstream analyses.

To accomplish the proposed standardized methods, we have released a set of version controlled open-sourced code to facilitate this process.

Recommendations

Input Data: When planning to construct a transcriptome, the first question to ponder is the type and quantity of data required. While this will be somewhat determined by the specific goals of the study and availability of tissues, there are some general guiding principals. As of 2014, Illumina continues to offer the most flexibility in terms of throughout, analytical tractability, and cost. It is worth noting however, that long-read (e.g. PacBio) transcriptome sequencing is just beginning to emerge as an alternative, particularly for researchers interested in understanding isoform complexity.

For the typical transcriptome study, one should plan to generate a reference based on 1 or more tissue types. From each tissue, one should be generating between 50M and 100M strand-specific paired-end reads. Read length should be at least 100bp, with longer reads aiding in isoform reconstruction and contiguity. Because sequence polymorphism increases the complexity of the *de bruijn* graph, and therefore may negatively effect the assembly itself, the reference transcriptome should be generated from reads corresponding to a single individual. When more then one individual is required to meet other requirements (e.g. number of reads), keeping the number of individuals to a minimum is paramount.

Quality Control of Sequence Read Data: Before assembly, it is critical that appropriate quality control steps are implemented. It is often helpful to generate some metrics of read quality on the raw data. Though this step may well be fairly unrepresentative of the true dataset quality, it is often informative and instructive. Several software packages are available– we are fond of SolexaQA and FastQC. These raw reads should be copied, compressed, and archived.

After visualizing the raw data, a vigorous adapter trimming step is implemented, typically using Trimmomatic. With adapter trimming may be a quality trimming step, though caution is required, as aggressive trimming may have detrimental effects on assembly quality. Specifically, we recommend trimming at Phred=2, a threshold associated with removal of the lowest quality bases. After adapter and quality trimming, it is recommended to once again visualize the data using SolexaQC. The .gz compressed reads are now ready for assembly.

Assembly: Assembly of transcriptome data is a ... Trinity is great, but is currently constrained to use a single kmer. In contrast, other assemblers (e.g. SOAPdenoveTrans) allows the used to select any value for k, which while increasing the time it takes to optimize assembly, may afford the ability to fine-tune the results, as well as implement a multi-kmer assembly approach.

Post-assembly transcriptome verification: Basically die N50, focus on functional metrics, transrate, etc..

Acknowledgments

References

1. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11: 485.

2. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30: 2114–2120.

3. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17: pp. 10–12.

4. Pell J, Pell J, Hintze A, Hintze A, Canino-Koning R, et al. (2012) Scaling metagenome sequence assembly with probabilistic *de Bruijn* graphs. Proceedings of the National Academy of Sciences 109: 13272–13277.

5. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 8: 1494–1512.

6. Xie Y, Wu G, Tang J, Luo R, Patterson J, et al. (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics (Oxford, England) 30: 1660–1666.

Figures

Tables