

The proper care and feeding of your transcriptome

Richard Smith-Unna¹, Matthew D MacManes³,

¹ University of Cambridge

³ Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham NH, USA

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

1 Abstract

2 Some abstract

3 Introduction

4 For biologists interested in understanding the relationship between fitness, genotype, and phe-
5 notype, modern sequencing technologies provide for an unprecedented opportunity to gain a
6 deep understanding of genome level processes that together, underlie adaptation. Transcrip-
7 tome sequencing has been particularly influential, and as a direct result, a diverse toolset for
8 the assembly and analysis of transcriptome exists. Notable amongst the wide array of tools
9 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
10 (3)), read normalization (khmer (4)), assembly (Trinity (5), SOAPdenovoTrans (6)) and assem-
11 bly verificaton (transrate <https://github.com/Blahah/transrate> and RSEM-eval (?)).
12

13

14 The ease with which these tools may be used to to produce transcriptome assemblies belies
15 that true complexity underlying the overall process. Indeed, the subtle (and not so subtle)
16 methodological challenges associated with transcriptome reconstruction means that you can
17 easily fuck it up. Amongst the most challenging include isoform reconstruction, simultaneous
18 assembly of low- and high-coverage transcripts, and [] (? ?), which together make good tran-
19 scriptome assembly really difficult.

20

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

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

28 Recommendations

29 INPUT DATA: When planning to construct a transcriptome, the first question to ponder is the
30 type and quantity of data required. While this will be somewhat determined by the specific goals
31 of the study and availability of tissues, there are some general guiding principals. As of 2014,
32 Illumina continues to offer the most flexibility in terms of throughout, analytical tractability,
33 and cost. It is worth noting however, that long-read (e.g. PacBio) transcriptome sequencing is
34 just beginning to emerge as an alternative (7), particularly for researchers interested in under-
35 standing isoform complexity.

36
37 For the typical transcriptome study, one should plan to generate a reference based on 1 or
38 more tissue types. From each tissue, one should be generating between 50M and 100M strand-
39 specific paired-end reads. Read length should be at least 100bp, with longer reads aiding in
40 isoform reconstruction and contiguity. Because sequence polymorphism increases the complex-
41 ity of the *de bruijn* graph, and therefore may negatively effect the assembly itself, the reference
42 transcriptome should be generated from reads corresponding to a single individual. When more
43 than one individual is required to meet other requirements (e.g. number of reads), keeping the
44 number of individuals to a minimum is paramount.

45
46 QUALITY CONTROL OF SEQUENCE READ DATA: Before assembly, it is critical that appro-
47 priate quality control steps are implemented. It is often helpful to generate some metrics of read
48 quality on the raw data. Though this step may well be fairly unrepresentative of the true dataset
49 quality, it is often informative and instructive. Several software packages are available— we are
50 fond of SolexaQA and FastQC. These raw reads should be copied, compressed, and archived.

51
52 After visualizing the raw data, a vigorous adapter trimming step is implemented, typically
53 using Trimmomatic. With adapter trimming may be a quality trimming step, though caution is
54 required, as aggressive trimming may have detrimental effects on assembly quality. Specifically,
55 we recommend trimming at Phred=2, a threshold associated with removal of the lowest quality
56 bases. After adapter and quality trimming, it is recommended to once again visualize the data
57 using SolexaQC. The .gz compressed reads are now ready for assembly.

58
59 ASSEMBLY: Assembly of transcriptome data is a ... Trinity is great, but is currently con-
60 strained to use a single kmer. In contrast, other assemblers (e.g. SOAPdenoveTrans) allows the
61 used to select any value for k, which while increasing the time it takes to optimize assembly, may
62 afford the ability to fine-tune the results, as well as implement a multi-kmer assembly approach.

63

64 POST-ASSEMBLY TRANSCRIPTOME VERIFICATION: Basically die N50, focus on functional
65 metrics, transrate, etc..
66

67 Acknowledgments

68 References

- 69 1. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illu-
70 mina second-generation sequencing data. *BMC Bioinformatics* 11: 485.
- 71 2. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina
72 sequence data. *Bioinformatics (Oxford, England)* 30: 2114–2120.
- 73 3. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing
74 reads. *EMBnetjournal* 17: pp. 10–12.
- 75 4. Pell J, Pell J, Hintze A, Hintze A, Canino-Koning R, et al. (2012) Scaling metagenome se-
76 quence assembly with probabilistic *de Bruijn* graphs. *Proceedings of the National Academy*
77 *of Sciences* 109: 13272–13277.
- 78 5. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) *De novo*
79 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
80 generation and analysis. *Nature protocols* 8: 1494–1512.
- 81 6. Xie Y, Wu G, Tang J, Luo R, Patterson J, et al. (2014) SOAPdenovo-Trans: *de novo*
82 transcriptome assembly with short RNA-Seq reads. *Bioinformatics (Oxford, England)* 30:
83 1660–1666.
- 84 7. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, et al. (2013) Characterization of
85 the human ESC transcriptome by hybrid sequencing. *PNAS* 110: 201320101–30.

86 Figures

87 Tables