Report: Quantifying the abundance of transcripts in Kallisto from transciptomes assembled in Trinity

Pete Rigas

Abstract

We implement Kallisto, an open source package developed by the Pachter Lab at Caltech for quantifying abundance of transcripts, given 26 input FASTA files generated by the Trinity transcriptome assembly pipeline. De novo, we identify the abundance of approximately $2,000,000$ full length transcripts, with Kalliso processing $1,330,514,396$ k-mers, of which $545,770,929$ are unique, from which $26,187$ unitigs are split to construct $6,007,667$ contigs across 26 de Bruijn graphs consisting of $480,822,288$ k-mers. After instantiating the index files containing more than $1,819,722$ reads, $1,817,831$ reads are pseudoaligned over 365 iterations of the Expectation-Maximization algorithm. We conclude by providing plots of different quantities which are approximated by Kallisto, including the transcripts per million, and corresponding length, and effective length, of each target. [1]

1 Introduction

1.1 Overview

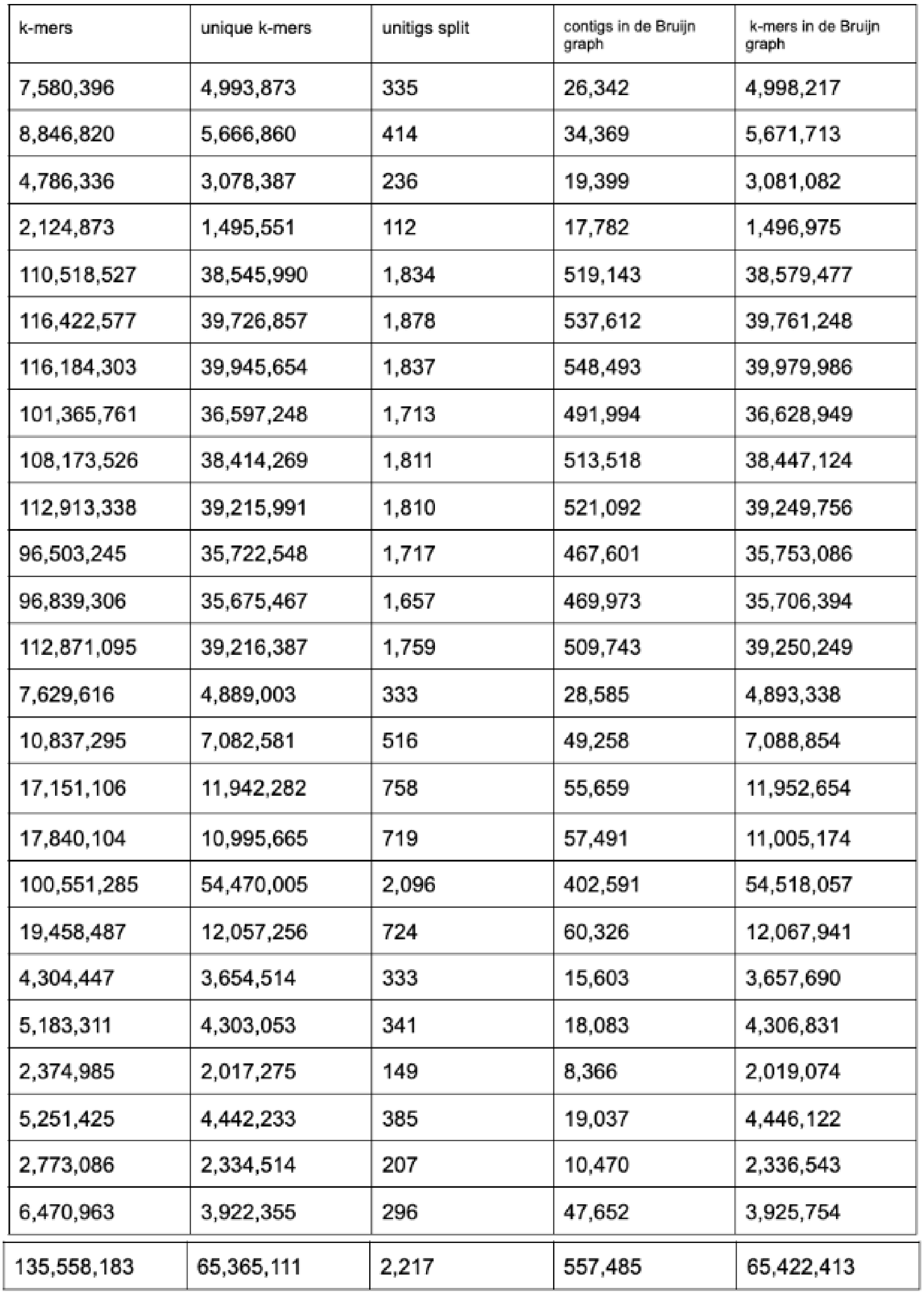
Transcriptome assembly pipelines, such as Trinity, have emerged as valuable computatioonal resources for analyzing bulk RNA sequencing for constructing full length transcripts without any reference to the genome, in which sequencing is processed in a procedure consisting of three steps: Inchworm, Chrysalis, and Butterfly [5,6]. Depending upon the complexity of sequencing gathered through the read length, the Trinity transcriptome assembly can easily have a runtime of several hours, in which RNA-seq data is processed from an input FASTA, or FASTQ, file, for alternatively splicing transcripts, after which de Bruijn graphs are constructed for clusters of Inchworm contigs, ultimately resulting in the assembly of full-length transcripts corresponding to paralogous genes. To further explore computational prospects at the forefront of probabilistic quantification [1], variational inference [3], differential expression [4], and various comparative analyses [2], we run Kallisto, an open source package developed by the Pachter Lab, on output FASTA files generated by the Trinity transcriptome assembly, in the hopes of quantifying abundance of hundreds of thousands of transcripts. In comparison to other bioinformatic pipelines, Kallisto relies upon pseudoalignment, which not only preserves key information required for the quantification of full-length transcripts, but also has a very short runtime, while simultaneously outperforming existing tools on several benchmarks [1].

The collection of full-length transcripts that we analyze from the Trinity transcriptome assembly consists of approximately $2,000,000$ genes. When creating the index files that are required for further downstream processes in Kallisto, we feed the algorithms single-end reads. Following the creation of the index files, which involve the construction of $6,007,667$ contigs from $26,187$ unitigs, for the ensuing quantification steps Kallisto creates equivalence classes for quantifying abundance, from which output TSV files are generated. From a fixed number of bootstrap samples, abundance estimates are generated in output HDF5 and TSV files, both of which include columns indicating measurements of the length, effective length, and transcripts per million (TPM), the last of which is related to the expected abundance of a transcript that one would observe given sequencing of one million full length transcripts. To conclude our analyses, we generate plots of the abundance profiles given each of the 26 output directories.

1.2 Paper organization

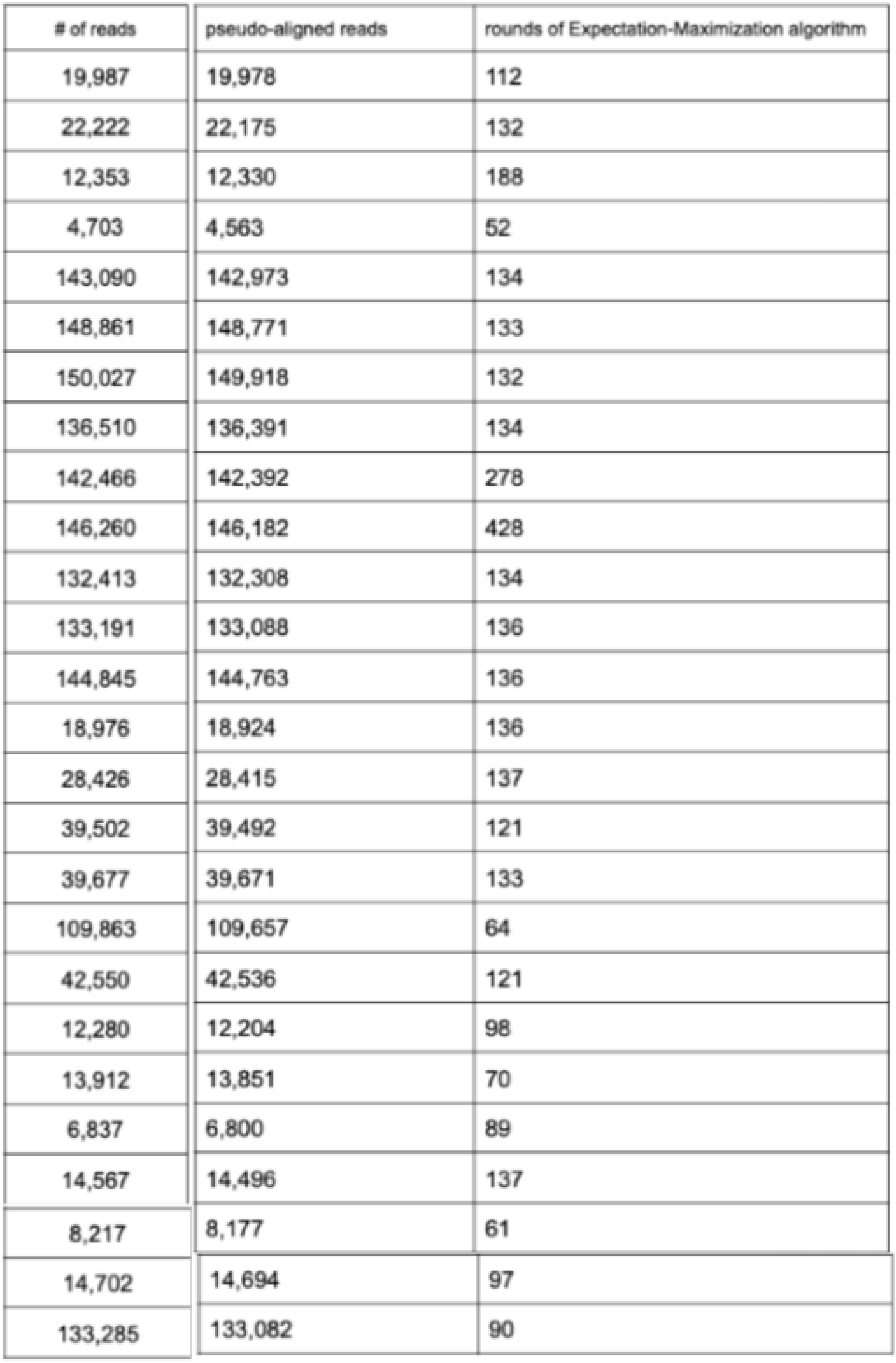
In the remaining pages, we provide tables outlining the number of k-mers processed when creating Kallisto index files, in addition to the number of unitigs split for constructing the de Bruijn graphs. From the total number of pseudoaligned reads, we graphically represent the approximate abundance of transcripts assembled from each Trinity compilation, in addition to the effective length versus length of different genes from targets produced by the Trinity transcriptome assembly.

[1] ***Keywords***: Kallisto, Trinity, transcriptome assembly, mRNA, transcript abundance



k-mers	unique k-mers	unitigs split	contigs in de Bruijn graph	k-mers in de Bruijn graph
7,580,396	4,993,873	335	26,342	4,998,217
8,846,820	5,666,860	414	34,369	5,671,713
4,786,336	3,078,387	236	19,399	3,081,082
2,124,873	1,495,551	112	17,782	1,496,975
110,518,527	38,545,990	1,834	519,143	38,579,477
116,422,577	39,726,857	1,878	537,612	39,761,248
116,184,303	39,945,654	1,837	548,493	39,979,986
101,365,761	36,597,248	1,713	491,994	36,628,949
108,173,526	38,414,269	1,811	513,518	38,447,124
112,913,338	39,215,991	1,810	521,092	39,249,756
96,503,245	35,722,548	1,717	467,601	35,753,086
96,839,306	35,675,467	1,657	469,973	35,706,394
112,871,095	39,216,387	1,759	509,743	39,250,249
7,629,616	4,889,003	333	28,585	4,893,338
10,837,295	7,082,581	516	49,258	7,088,854
17,151,106	11,942,282	758	55,659	11,952,654
17,840,104	10,995,665	719	57,491	11,005,174
100,551,285	54,470,005	2,096	402,591	54,518,057
19,458,487	12,057,256	724	60,326	12,067,941
4,304,447	3,654,514	333	15,603	3,657,690
5,183,311	4,303,053	341	18,083	4,306,831
2,374,985	2,017,275	149	8,366	2,019,074
5,251,425	4,442,233	385	19,037	4,446,122
2,773,086	2,334,514	207	10,470	2,336,543
6,470,963	3,922,355	296	47,652	3,925,754
135,558,183	65,365,111	2,217	557,485	65,422,413

Figure 1: *Performance benchmarks for creating Kallisto indices.* 480, 822, 288 k-mers are instantiated in 26 de Bruijn graphs.



# of reads	pseudo-aligned reads	rounds of Expectation-Maximization algorithm
19,987	19,978	112
22,222	22,175	132
12,353	12,330	188
4,703	4,563	52
143,090	142,973	134
148,861	148,771	133
150,027	149,918	132
136,510	136,391	134
142,466	142,392	278
146,260	146,182	428
132,413	132,308	134
133,191	133,088	136
144,845	144,763	136
18,976	18,924	136
28,426	28,415	137
39,502	39,492	121
39,677	39,671	133
109,863	109,657	64
42,550	42,536	121
12,280	12,204	98
13,912	13,851	70
6,837	6,800	89
14,567	14,496	137
8,217	8,177	61
14,702	14,694	97
133,285	133,082	90

Figure 2: *Performance Benchmarks for Kallisto quantification.* Of the $1,819,722$ reads that are processed, $1,817,831$ reads are pseudoaligned over 365 rounds of the E-M algorithm.

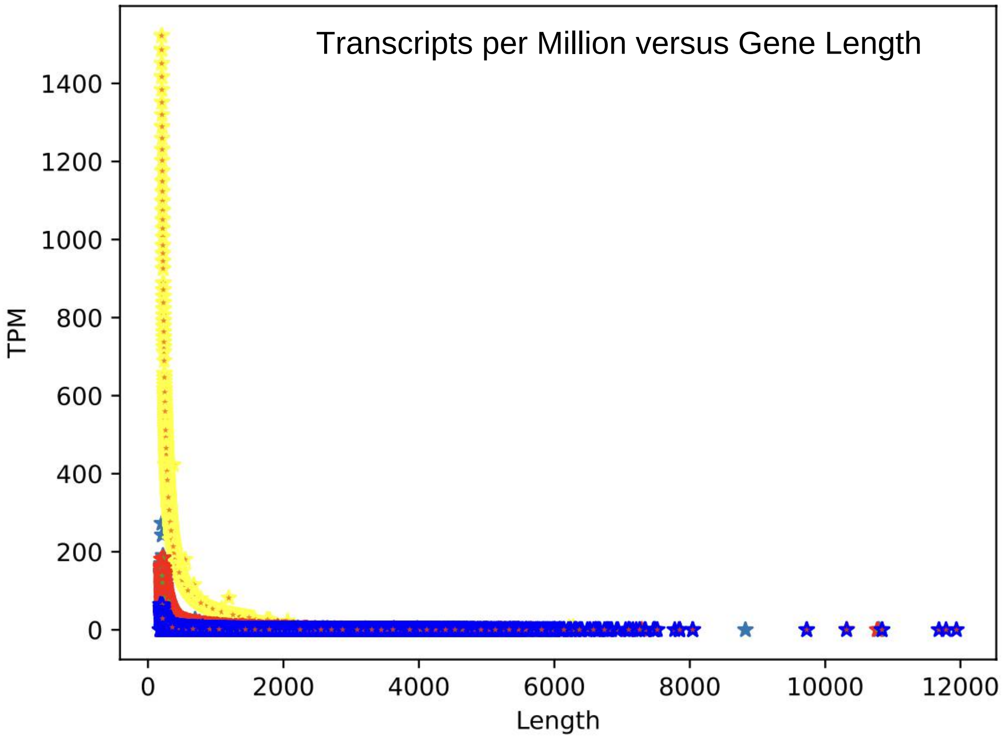


Figure 3: *Plotting the transcripts per million versus gene length for* 3 *iterations of Kallisto on transcriptomes assembled in Trinity.* The plot above exhibits how the relationship between TPM and gene length can differ depending upon the assembled transcriptome that is provided as input to Kallisto, in which the points plotted in yellow decay more slowly with respect to the gene length than do points plotted in blue and red.

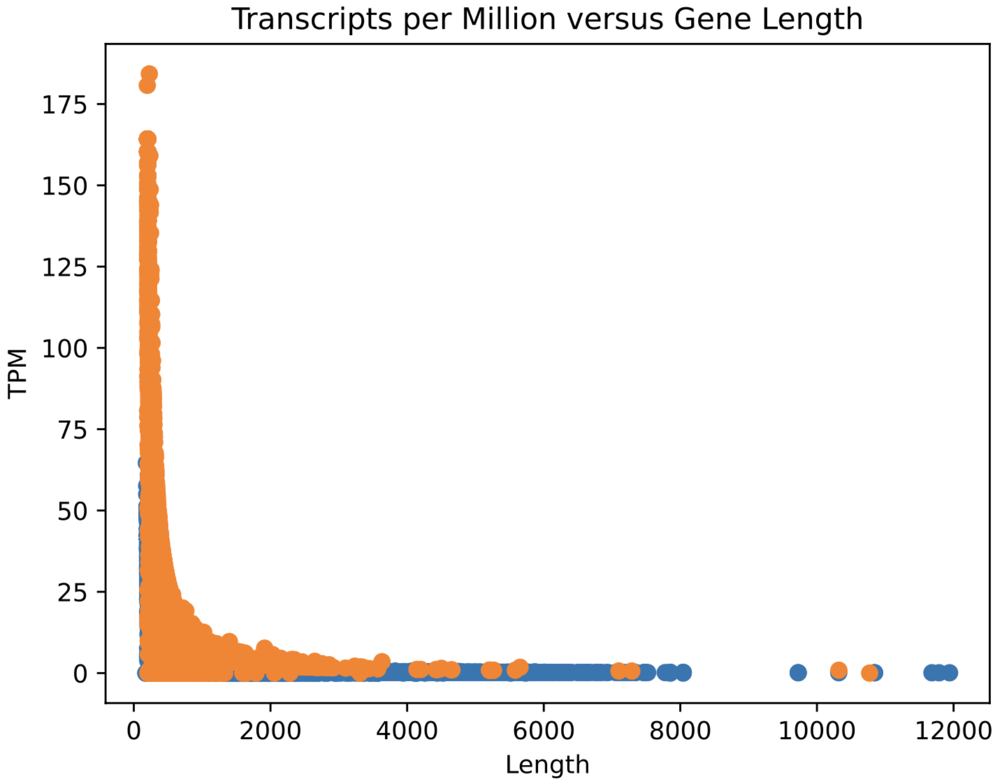


Figure 4: *Plotting the transcripts per million versus gene length for* 2 *iterations of Kallisto on transcriptomes assembled in Trinity.*

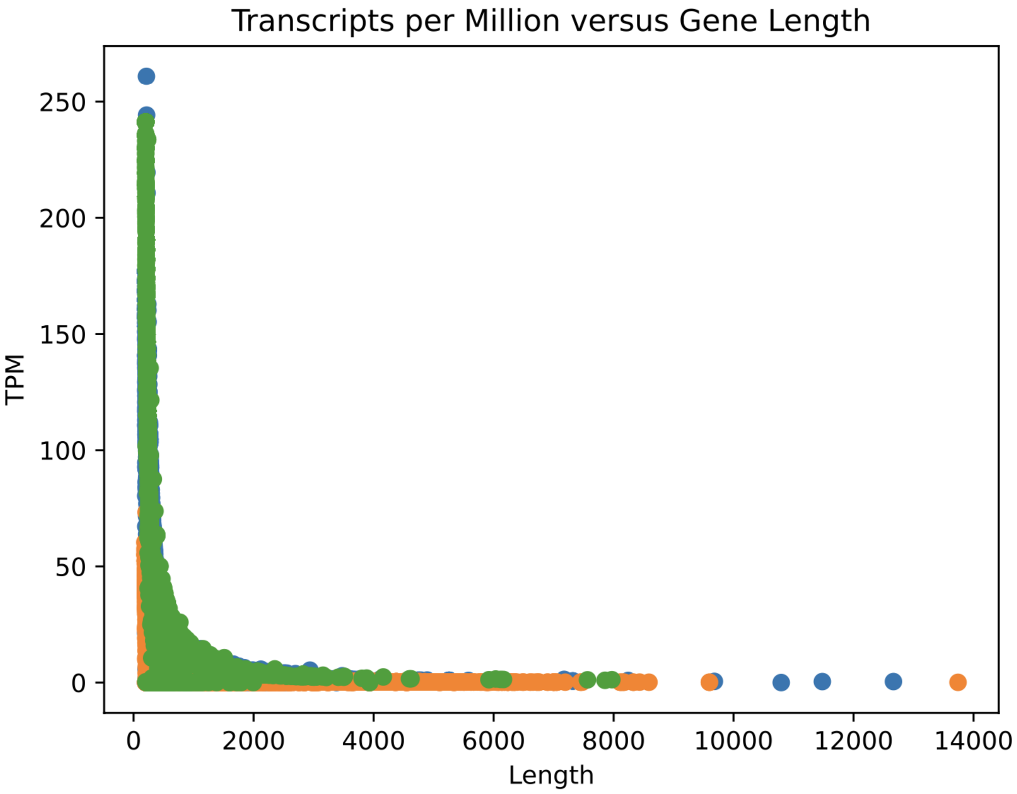


Figure 5: *Plotting the transcripts per million versus gene length for* 2 *iterations of Kallisto on transcriptomes assembled in Trinity.*

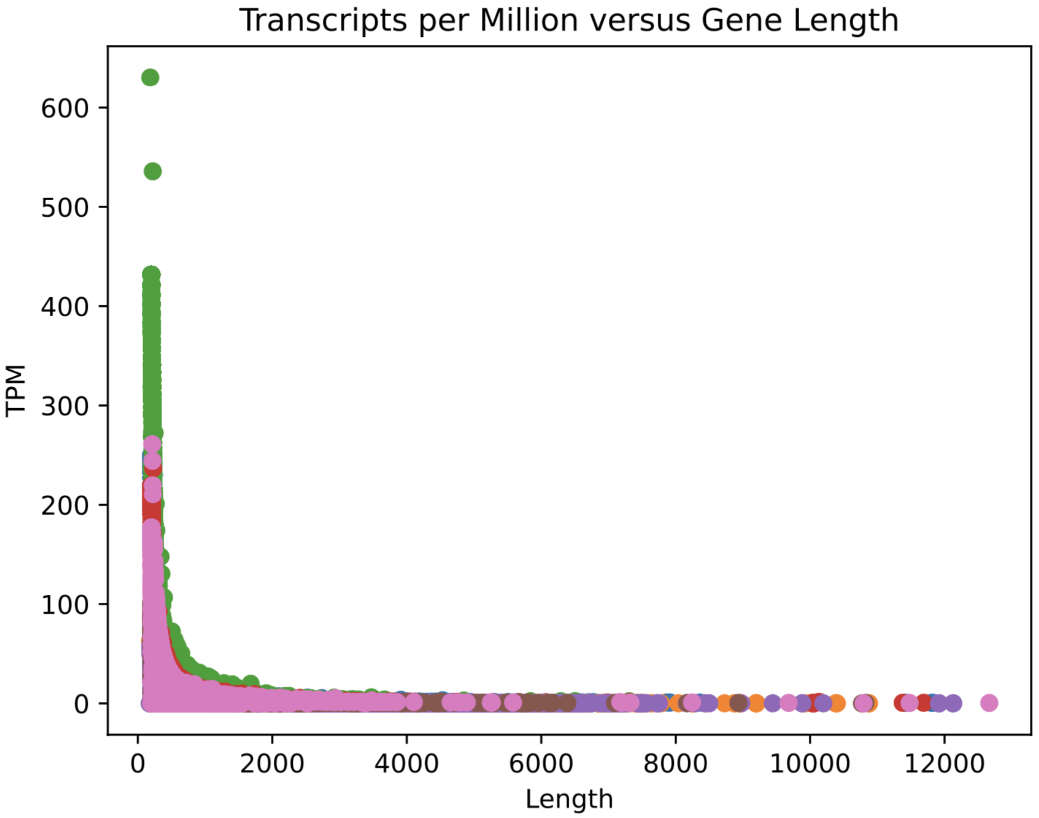


Figure 6: *Plotting the transcripts per million versus gene length for* 4 *iterations of Kallisto on transcriptomes assembled in Trinity.*

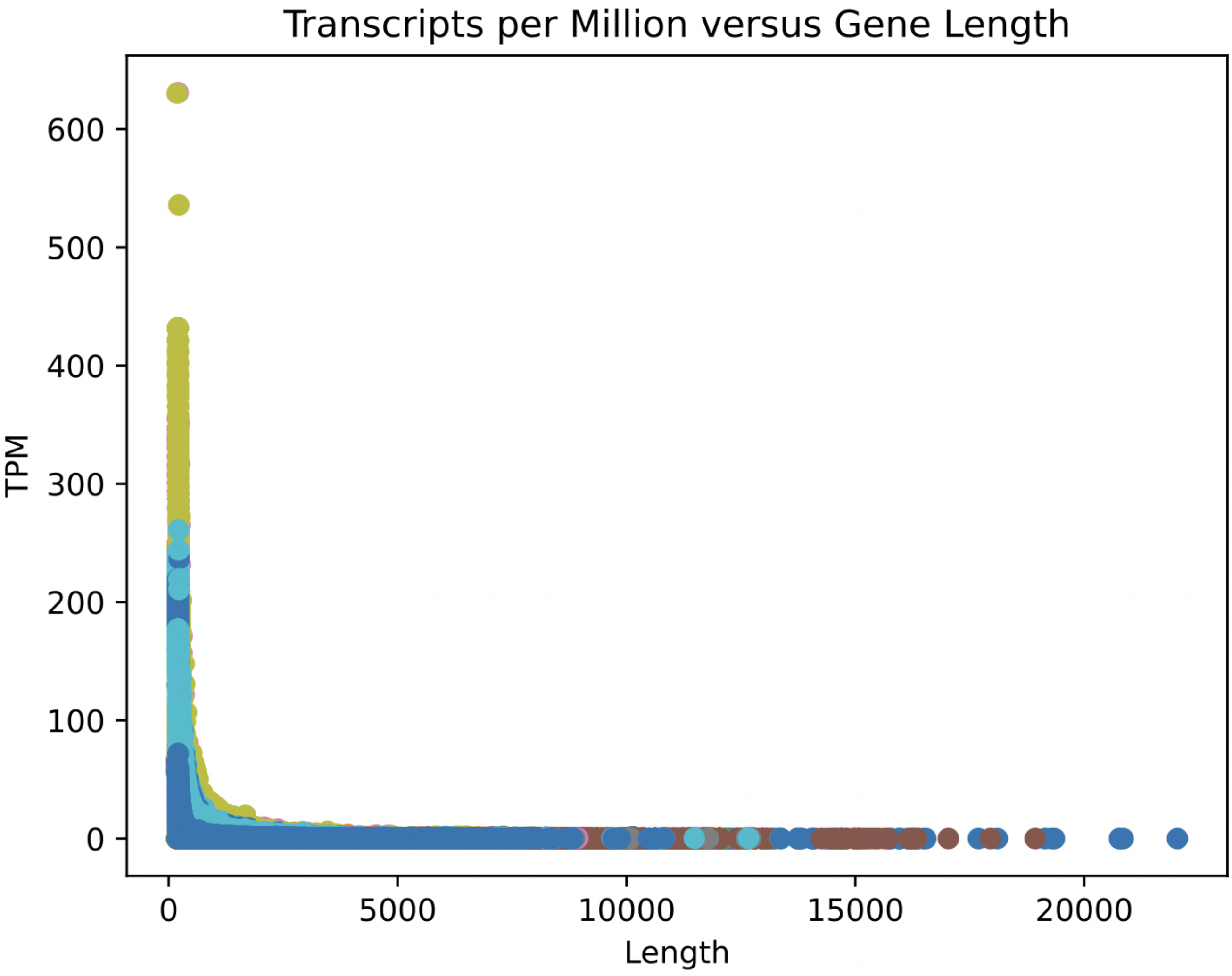


Figure 7: *Plotting the transcripts per million versus gene length for* 20 *iterations of Kallisto on transcriptomes assembled in Trinity.* In the plot above, the TPM, which is approximated by Kallisto as the abundance, is plotted against the length of each gene. From data points included in the brighter orange color, one observes that the genes with the highest TPM value are typically of shorter length, while other data points that are included in blue and red indicate that genes with a longer length have a lower TPM value. In comparison to previous plots of the TPM versus gene length that are provided in *Figure 3* - *Figure 6*, the plot above demonstrates how the transcript abundance is expected to change with respect to the gene length for different transcriptome data sets.

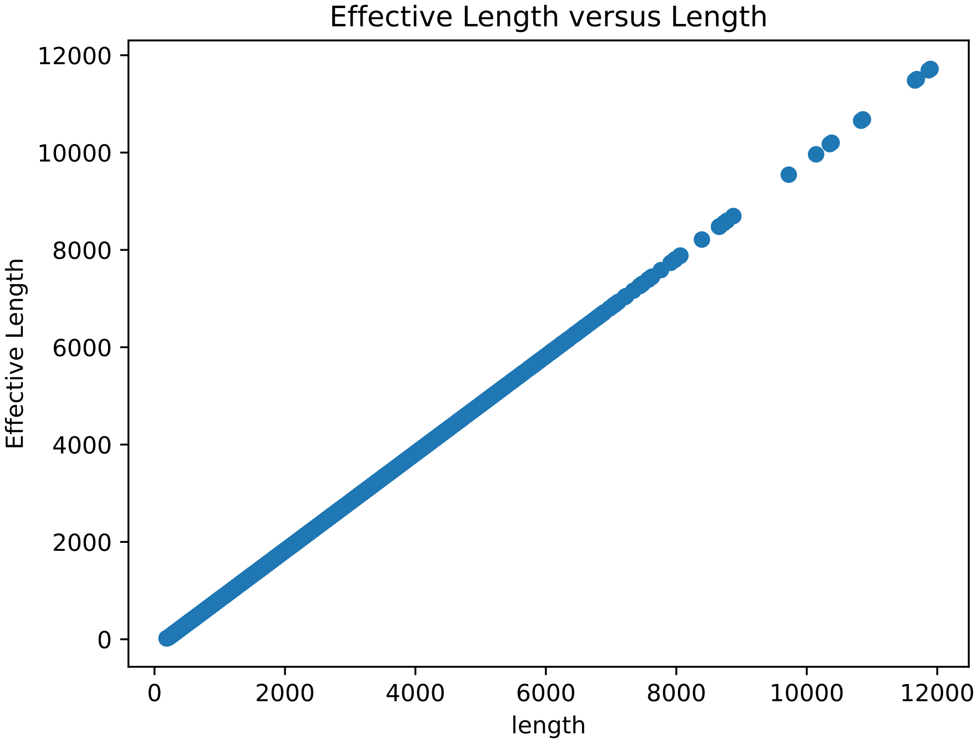


Figure 8: *Plotting the Effective length versus length for one transcriptome assembled in Trinity.*

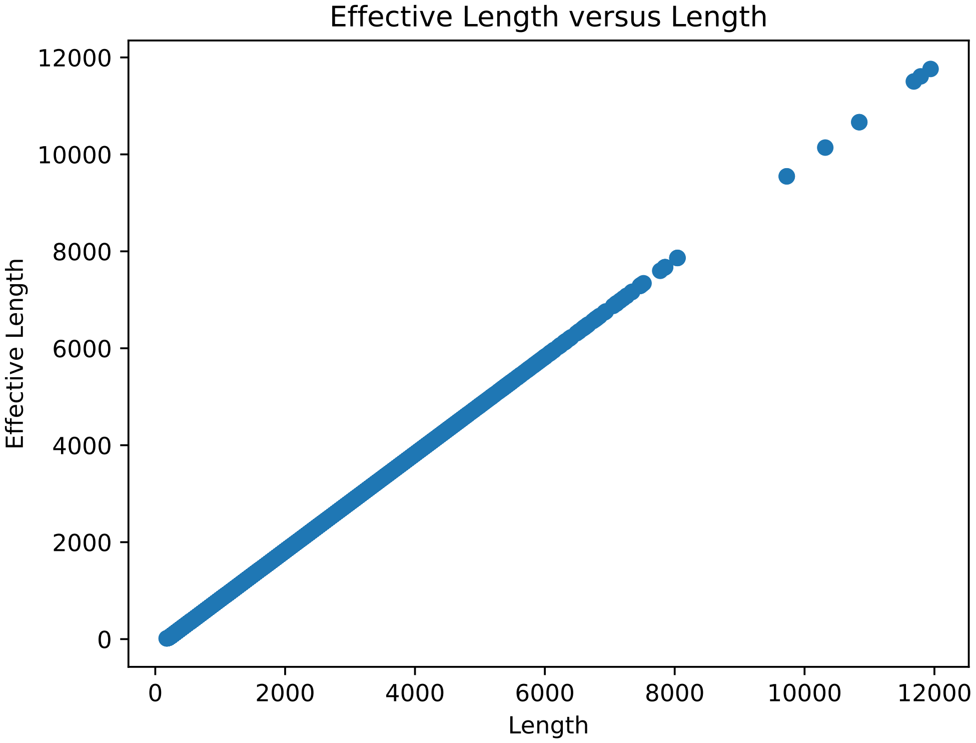


Figure 9: *Plotting the Effective length versus length for one transcriptome assembled in Trinity.*

1.3 Data availability

An example of the output produced by Kallisto for one of the FASTA files is available at https : //github.com/peter − beep/Kallisto. The remaining 25 output directories are available upon request.

2 References

[1] Bray, N., Pimentel, H., Melsted, P. et al. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol34, 525–527 (2016). https://doi.org/10.1038/nbt.3519

[2] Clarke, K., Yang, Y., Marsh, R. et al. Comparative analysis of de novo transcriptome assembly. Sci. China Life Sci. 56, 156–162 (2013). https://doi.org/10.1007/s11427-013-4444-x

[3] Ferreira, P.F., Carvalho, A.M., Vinga, S. (2020). Variational Inference in Probabilistic Single-cell RNA-seq Models. In: Raposo, M., Ribeiro, P., Sério, S., Staiano, A., Ciaramella, A. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2018. Lecture Notes in Computer Science, vol 11925. Springer, Cham. https : //doi.org/10.1007/978 − 3 − 030 − 34585 − 32

[4] Peter Glaus and others, Identifying differentially expressed transcripts from RNA-seq data with biological variation, Bioinformatics, Volume 28, Issue 13, July 2012, Pages 1721–1728, https : //doi.org/10.1093/bioinformat

[5] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52. doi: 10.1038/nbt.1883. PubMed PMID: 21572440

[6] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494-512. Open Access in PMC doi: 10.1038/nprot.2013.084. Epub 2013 Jul 11. PubMed PMID:23845962

[7] Henschel R, Lieber M, Wu L, Nista, PM, Haas BJ, LeDuc R. Trinity RNA-Seq assembler performance optimization. XSEDE 2012 Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond. ISBN: 978-1-4503-1602-6 doi: 10.1145/2335755.2335842

[8] Wang,S., Gribskov, M. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis, Bioinformatics, Volume 33, Issue 3, February 2017, Pages 327–333, https://doi.org/10.1093/bioinformatics/btw625

[9] Zhang, A.W., O'Flanagan, C., Chavez, E.A. et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods 16, 1007–1015 (2019). https : //doi.org/10.1038/ s41592 − 019 − 0529 − 1