The Oyster River Protocol: A Multi Assembler and Kmer Approach For *de novo* Transcriptome Assembly

Matthew D. MacManes[1, 2, *, •, ⋆]

[1] Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham NH, USA

[2] Hubbard Center For Genomic Studies

* E-mail: macmanes@gmail.com

• Twitter: @MacManes

⋆ Mailing Address: ~~46 College Road, 189 Rudman~~ 35 Colovos Road, 434 Gregg Hall. Durham NH 03824

Abstract

Characterizing transcriptomes in non-model organisms has resulted in a massive increase in the understanding of biological phenomena. This boon, largely made possible via high-throughput sequencing, means that studies of functional, evolutionary and population genomics are now being done by hundreds or even thousands of labs around the world. For many, these studies begin with a *de novo* transcriptome assembly, which is a technically complicated process involving several discrete steps. The Oyster River Protocol (ORP), described here, implements a standardized and benchmarked set of bioinformatic processes, resulting in an assembly with enhanced qualities over other standard assembly methods. Specifically, ORP produced assemblies have higher `Detonate` and `TransRate` scores and mapping rates, which is largely a product of the fact that it leverages a multi-assembler and kmer assembly process, thereby bypassing the shortcomings of any one approach. These improvements are important, as previously unassembled transcripts are included in ORP assemblies, resulting in a significant enhancement of the power of downstream analysis. Further, as part of this study, I show that assembly quality is unrelated with the number of reads generated, above 30 million reads. **Code Availability:** The version controlled open-source code is available at `https://github.com/macmanes-lab/Oyster_River_Protocol`. Instructions for software installation and use, and other details are available at `http://oyster-river-protocol.rtfd.org/`.

1 Introduction

For all biology, modern sequencing technologies have provided for an unprecedented opportunity to gain a deep understanding of genome level processes that underlie a very wide array of natural phenomena, from intracellular metabolic processes to global patterns of population variability. Transcriptome sequencing has been influential (Mortazavi, Williams, Mccue, Schaeffer, and Wold, 2008; Z. Wang, Z. Wang, Gerstein, and Snyder, 2009), particularly in functional genomics (Lappalainen et al., 2013; Cahoy et al., 2008), and has resulted in discoveries not possible even just a few years ago. This in large part is due to the scale at which these studies may be conducted (X. Li et al., 2017; Tan et al., 2017). Unlike studies of adaptation based on one or a small number of candidate genes (*e.g.,* (Fitzpatrick, Ben-Shahar, Vet, Smid, G. E. Robinson, and Sokolowski, 2005; Panhuis, 2006)), modern studies may assay the entire suite of expressed transcripts – the transcriptome – simultaneously. In addition to issues of scale, as a direct result of enhanced dynamic range, newer sequencing studies have increased ability to simultaneously reconstruct and quantitate lowly- and highly-expressed transcripts (Wolf, 2013; Vijay, Poelstra, Künstner, and Wolf, 2013). Lastly, improved methods

for the detection of differences in gene expression (*e.g.,* (M. D. Robinson, McCarthy, and Smyth, 2010; Love, Huber, and anders, 2014)) across experimental treatments have resulted in increased resolution for studies aimed at understanding changes in gene expression.

As a direct result of their widespread popularity, a diverse toolset for the assembly of transcriptome exists, with each potentially reconstructing transcripts others fail to reconstruct. Amongst the earliest of specialized *de novo* transcriptome assemblers were the packages `Trans-ABySS` (Robertson et al., 2010), `Oases` (Marcel H Schulz, Daniel R Zerbino, Vingron, and Ewan Birney, 2012), and `SOAPdenovoTrans` (Xie et al., 2014), which were fundamentally based on the popular *de Bruijn* graph-based genome assemblers `ABySS` (Simpson, Wong, S D Jackman, Schein, Jones, and Birol, 2009), `Velvet` (D R Zerbino and Birney, 2008), and `SOAP` ~~R. Li, Y. Li, Kristiansen, and J. Wang, 2008~~ (R. Li, Y. Li, Kristiansen, and J. Wang, 2008) respectively. These early efforts gave rise to a series of more specialized *de novo* transcriptome assemblers, namely `Trinity` (Haas et al., 2013), and `IDBA-Tran` (Peng et al., 2013). While the *de Bruijn* graph approach remains powerful, newly developed software explores novel parts of the algorithmic landscape, offering substantial benefits, assuming novel methods reconstruct different fractions of the transcriptome. `BinPacker` (J. Liu, G. Li, Chang, Yu, B. Liu, McMullen, Chen, and Huang, 2016), for instance, abandons the *de Bruijn* graph approach to model the assembly problem after the classical bin packing problem, while `Shannon` (Kannan, Hui, Mazooji, Pachter, and Tse, 2016) uses information theory, rather than a set of software engineer-decided heuristics. These newer assemblers, by implementing fundamentally different assembly algorithms, may reconstruct fractions of the transcriptome that other assemblers fail to accurately assemble.

In addition to the variety of tools available for the *de novo* assembly of transcripts, several tools are available for pre-processing of reads via read trimming ((*e.g.,* `Skewer` (Jiang, Lei, Ding, and Zhu, 2014), `Trimmomatic` (Bolger, Lohse, and Usadel, 2014), `Cutadapt` ~~M. Martin, 2011~~(M. Martin, 2011)), read normalization (`khmer` (Pell, Hintze, Canino-Koning, Howe, Tiedje, and Brown, 2012)), and read error correction (`SEECER` (Le, M H Schulz, McCauley, Hinman, and Bar-Joseph, 2013) and `RCorrector` (Song and Florea, 2015), `Reptile` ~~X. Yang, X. Yang, Dorman, Dorman, Aluru, and Aluru, 2010~~(X. Yang, X. Yang, Dorman, Dorman, Aluru, and Aluru, 2010)). Similarly, benchmarking tools that evaluate the quality of assembled transcriptomes including `TransRate` (R. Smith-Unna, Boursnell, Patro, Hibberd, and Kelly, 2016), `BUSCO` (Benchmarking Universal Single-Copy Orthologs - (Simão, Waterhouse, Ioannidis, Kriventseva, and Zdobnov, 2015)), and `Detonate` (B. Li, Fillmore, Bai, Collins, Thomson, Stewart, and Dewey, 2014) have been developed. Despite the development of these evaluative tools, this manuscript describes the first systematic effort coupling them with the development of a *de novo* transcriptome assembly pipeline.

The ease with which these tools may be used to produce and characterize transcriptome assemblies belies the true complexity underlying the overall process (Ungaro, Pech, J.-F. Martin, McCairns, Mévy, Chappaz, and Gilles, 2017; S. Wang and Gribskov, 2017; Moreton, Izquierdo, and Emes, 2015; Y. Yang and Smith, 2013). Indeed, the subtle (and not so subtle) methodological challenges associated with transcriptome reconstruction may result in highly variable assembly quality. In particular, while most tools run using default settings, these defaults may be sensible only for one specific (often unspecified) use case or data type. Because parameter optimization is both dataset-dependent and factorial in nature, an exhaustive optimization particularly of entire pipelines, is never possible. Given this, the production of a *de novo* transcriptome assembly requires a large investment in time and resources, with each step requiring careful consideration. Here, I propose an evidence-based protocol for assembly that results in the production of high quality transcriptome assemblies, across a variety of commonplace experimental conditions or taxonomic groups.

This manuscript describes the development of The Oyster River Protocol[1] for transcriptome assembly. It explicitly considers and attempts to address many of the shortcomings described in (Vijay, Poelstra, Künstner, and Wolf, 2013), by leveraging a multi-kmer and multi-assembler strategy. This innovation is critical, as all assembly solutions treat the sequence read data in ways that bias transcript recovery. Specifically, with the development of assembly software comes the use of a set of heuristics that are necessary given the scope of the assembly problem itself. Given each software development team carries with it a unique set of ideas related to these heuristics while implementing various assembly algorithms, individual assemblers exhibit unique assembly behavior. By leveraging a multi-assembler approach, the strengths of one assembler may complement the weaknesses of another. In addition to biases related to assembly heuristics, it is well known that assembly kmer-length has important effects on transcript reconstruction, with shorter kmers more efficiently reconstructing lower-abundance transcripts relative to more highly abundant transcripts. Given this, assembling with multiple different kmer lengths, then merging the resultant assemblies may effectively reduce this type of bias. Recognizing these issue, I hypothesize that an assembly that results from the combination of multiple different assemblers and lengths of assembly-kmers will be better than each individual assembly, across a variety of metrics.

In addition to developing an enhanced pipeline, the work suggests an exhaustive way of characterizing assemblies while making available a set of fully-benchmarked reference assemblies that may be used by other researchers in developing new assembly algorithms and pipelines. Although many other researchers have published comparisons of assembly methods, up until now these have been limited to single datasets assembled a few different ways (Marchant, Mougel, Mendonça, Quartier, Jacquin-Joly, Rosa, Petit, and Harry,

2016; Finseth and Harrison, 2014), thereby failing to provide more general insights.

2 Methods

2.1 Datasets

In an effort at benchmarking the assembly and merging protocols, I downloaded a set of publicly available RNAseq datasets (Table 1) that had been produced on the Illumina sequencing platform. These datasets were chosen to represent a variety of taxonomic groups, so as to demonstrate the broad utility of the developed methods. Because datasets were selected randomly with respect to sequencing center and read number, they are likely to represent the typical quality of Illumina data circa 2014-2017.

Table 1

Type	Accession	Species	Num. Reads	Read Length
Animalia	ERR489297	*Anopheles gambiae*	206M	100bp
Animalia	DRR030368	*Echinococcus multilocularis*	73M	100bp
Animalia	ERR1016675	*Heterorhabditis indica*	51M	100bp
Animalia	SRR2086412	*Mus musculus*	54M	100bp
Animalia	DRR036858	*Mus musculus*	114M	100bp
Animalia	DRR046632	*Oncorhynchus mykiss*	82M	76bp
Animalia	SRR1789336	*Oryctolagus cuniculus*	31M	100bp
Animalia	SRR2016923	*Phyllodoce medipapillata*	86M	100bp
Animalia	ERR1674585	*Schistosoma mansoni*	39M	100bp
Plant	DRR082659	*Aeginetia indica*	69M	90bp
Plant	DRR053698	*Cephalotus follicularis*	126M	90bp
Plant	DRR069093	*Hevea brasiliensis*	103M	100bp
Plant	SRR3499127	*Nicotiana tabacum*	30M	150bp
Plant	DRR031870	*Vigna angularis*	60M	100bp
Protozoa	ERR058009	*Entamoeba histolytica*	68M	100bp

Table 1 lists the datasets used in this study. All datasets are publicly available for download by accession number at the European Nucleotide Archive or NCBI Short Read Archive.

[1]Named the Oyster River Protocol because the ideas, and some of the code, was developed while overlooking the Oyster River, located in Durham, New Hampshire. NB, the naming assembly of protocols after bodies of water was, to the best of my knowledge, first done by C. Titus Brown (The Eel Pond Protocol: `http://khmer-protocols.readthedocs.io/en/latest/mrnaseq/index.html`), and may have subconsciously influenced me in naming this protocol.

2.2 Software

The Oyster River Protocol can be installed on the Linux platform, and does not require superuser privileges, assuming `Linuxbrew` (Shaun D Jackman and Inanç Birol, 2016) is installed. The software is implemented as a stand-alone makefile which coordinates all steps described below. All scripts are available at `https://github.com/macmanes-lab/Oyster_River_Protocol`, and run on the Linux platform. The software is version controlled and openly-licensed to promote sharing and reuse. A guide for users is available at `http://oyster-river-protocol.rtfd.io`.

2.3 Pre-assembly procedures

For all assemblies performed, Illumina sequencing adapters were removed from both ends of the sequencing reads, as were nucleotides with quality Phred ≤ 2, using the program `Trimmomatic` version 0.36 ~~Bolger, Lohse, and Usadel, 2014~~(Bolger, Lohse, and Usadel, 2014), following the recommendations from ~~Matthew D MacManes, 2014~~(Matthew D MacManes, 2014). After trimming, reads were error corrected using the software `RCorrector` version 1.0.2 ~~Song and Florea, 2015~~(Song and Florea, 2015), following recommendations from ~~Matthew David MacManes and Eisen, 2013~~(Matthew David MacManes and Eisen, 2013). The code for running this step of the Oyster River protocols is available at `https://github.com/macmanes-lab/Oyster_River_Protocol/blob/master/oyster.mk#L134`. The trimmed and error corrected reads were then subjected to *de novo* assembly.

2.4 Assembly

I assembled each trimmed and error corrected dataset using three different *de novo* transcriptome assemblers and three different kmer lengths, producing 4 unique assemblies. First, I assembled the reads using `Trinity` release 2.4.0 (Haas et al., 2013), and default settings (k=25), without read normalization. The decision to forgo normalization is based on previous work (Matthew D MacManes, 2015) showing slightly worse performance of normalized datasets. Next, the `SPAdes` RNAseq assembler (version 3.10) ~~Chikhi and Medvedev, 2014~~ (Chikhi and Medvedev, 2014) was used, in two distinct runs, using kmer sizes 55 and 75. Lastly, reads were assembled using the assembler `Shannon` version 0.0.2 ~~Kannan, Hui, Mazooji, Pachter, and Tse, 2016~~(Kannan, Hui, Mazooji, Pachter, and Tse, 2016), using a kmer length of 75. These assemblers were chosen based on the fact that they [1] use an open-science

development model, whereby end-users may contribute code, [2] are all actively maintained and are undergoing continuous development, and [3] occupy different parts of the algorithmic landscape.

This assembly process resulted in the production of four distinct assemblies. The code for running this step of the Oyster River protocols is available at `https://github.com/macmanes-lab/Oyster_River_Protocol/blob/master/oyster.mk#L142`.

2.5 Assembly Merging via OrthoFuse

To merge the four assemblies produced as part of the Oyster River Protocol, I developed new software that effectively merges transcriptome assemblies. Described in brief, `OrthoFuse` begins by concatenating all assemblies together, then forms groups of transcripts by running a version of `OrthoFinder` ~~Emms and Kelly, 2015~~ (Emms and Kelly, 2015) packaged with the ORP, modified to accept nucleotide sequences from the merged assembly. These groupings represent groups of homologous transcripts. While isoform reconstruction using short-read data is notoriously poor, by increasing the inflation parameter by default to I=4, it attempts to prevent the collapsing of transcript isoforms into single groups. After `Orthofinder` has completed, a modified version of `TransRate` version 1.0.3 ~~R. Smith-Unna, Boursnell, Patro, Hibberd, and Kelly, 2016~~ (R. Smith-Unna, Boursnell, Patro, Hibberd, and Kelly, 2016) which is packaged with the ORP, is run on the merged assembly, after which the best (= highest contig score) transcript is selected from each group and placed in a new assembly file to represent the entire group. The resultant file, which contains the highest scoring contig for each orthogroup, may be used for all downstream analyses. `OrthoFuse` is run automatically as part of the Oyster River Protocol, and additionally is available as a stand alone script, `https://github.com/macmanes-lab/Oyster_River_Protocol/blob/master/orthofuser.mk`.

2.6 Assembly Evaluation

All assemblies were evaluated using `ORP-TransRate`, `Detonate` version 1.11 ~~B. Li, Fillmore, Bai, Collins, Thomson, Stewart, and Dewey, 2014~~(B. Li, Fillmore, Bai, Collins, Thomson, Stewart, and Dewey, 201 `shmlast` version 1.2 (Scott, 2017), and `BUSCO` version 3.0.2 (Simão, Waterhouse, Ioannidis, Kriventseva, and Zdobnov, 2015). `TransRate` evaluates transcriptome assembly contiguity by producing a score based on length-based and mapping metrics, while `Detonate` conducts an orthogonal analysis, producing a score that is maximized by an assembly that is representative of input sequence read data. `BUSCO` evaluates assembly content by searching the assem-

blies for conserved single copy orthologs found in all Eukaryotes. I report default `BUSCO` metrics as described in Simão, Waterhouse, Ioannidis, Kriventseva, and Zdobnov, 2015(Simão, Waterhouse, Ioannidis, Kriventseva, and Zdobnov, 201 Specifically, "complete orthologs", are defined as query transcripts that are within 2 standard deviations of the length of the `BUSCO` group mean, while contigs falling short of this metric are listed as "fragmented". `Shmlast` implements the conditional reciprocal best hits (CRBH) test (Aubry, Kelly, Kümpers, R. D. Smith-Unna, and Hibberd, 2014), conducted in this case against the Swiss-Prot protein database (downloaded October, 2017) using an e-value of 1E-10.

In addition to the generation of metrics to evaluation the quality of transcriptome assemblies, I generated a distance matrix of assemblies for each dataset using the `sourmash` package (Titus Brown and Irber, 2016), in an attempt at characterizing the algorithmic landscape of assemblers. Specifically, each assembly was characterized using the `compute` function using 5000 independent sketches. The distance between assemblies was calculated using the `compare` function and a kmer length of 51. These distance matrices were visualized using the `isoMDS` function of the MASS package (`https://CRAN.R-project.org/package=MASS`).

2.7 Statistics

All statistical analyses were conducted in R version 3.4.0 (R Core Development Team, 2011). Violin plots were constructed using the beanplot (Kampstra, 2008) and the beeswarm R packages (`https://CRAN.R-project.org/package=beeswarm`). Expression distributions were plotted using the ggjoy package (`https://CRAN.R-project.org/package=ggjoy`).

3 Results and Discussion

Fifteen RNAseq datasets, ranging in size from (30-206M paired end reads) were assembled using the Oyster River Protocol and with `Trinity`. Each assembly was evaluated using the software `BUSCO`, `shmlast`, `Detonate`, and `TransRate`. From these, several metrics were chosen to represent the quality of the produced assemblies. Of note, all the assemblies produced as part of this work are available at `https://www.dropbox.com/sh/ehxvd0ont9ge8id/AABZxRCwcpaxb7rXWclTBbJga`, and will be moved to dataDryad after acceptance. A file containing the evaluative metrics is available at `https://github.com/macmanes-lab/Oyster_River_Protocol/blob/master/manuscript/orp.csv`, while the distance matrices are available within the folder

https://github.com/macmanes-lab/Oyster_River_Protocol/blob/master/manuscript/. R code used to conduct analyses and make figures is found at https://github.com/macmanes-lab/Oyster_River_Protocol/blob/master/manuscript/R-analysis.Rmd.

3.1 Assembled transcriptomes

The Trinity assembly of trimmed and error corrected reads generally completed on a standard Linux server using 24 cores, in less than 24 hours. RAM requirement is estimated to be close to 0.5Gb per million paired-end reads. The assemblies on average contained 176k transcripts (range 19k - 643k) and 97Mb (range 14MB - 198Mb). Other quality metrics will be discussed below, specifically in relation to the ORP produced assemblies.

ORP assemblies generally completed on a standard Linux server using 24 cores in three days. Typically Trinity was the longest running assembler, with the individual SPAdes assemblies being the shortest. RAM requirement is estimated to be 1.5Gb - 2Gb per million paired-end reads, with SPAdes requiring the most. The assemblies on average contained 153k transcripts (range 23k - 625k) and 64Mb (range 8MB - 181Mb).

The distance between assemblies of a given dataset were calculated using sourmash, and a MDS plot was generated (Figure 1). Interestingly, each assembler tends to produce a specific signature which is relatively consistent between the fifteen datasets. Shannon differentiates itself from the other assemblers on the first (x) MDS axis, while the other assemblers (SPAdes and Trinity) are separated on the second (y) MDS axis.

Figure 1

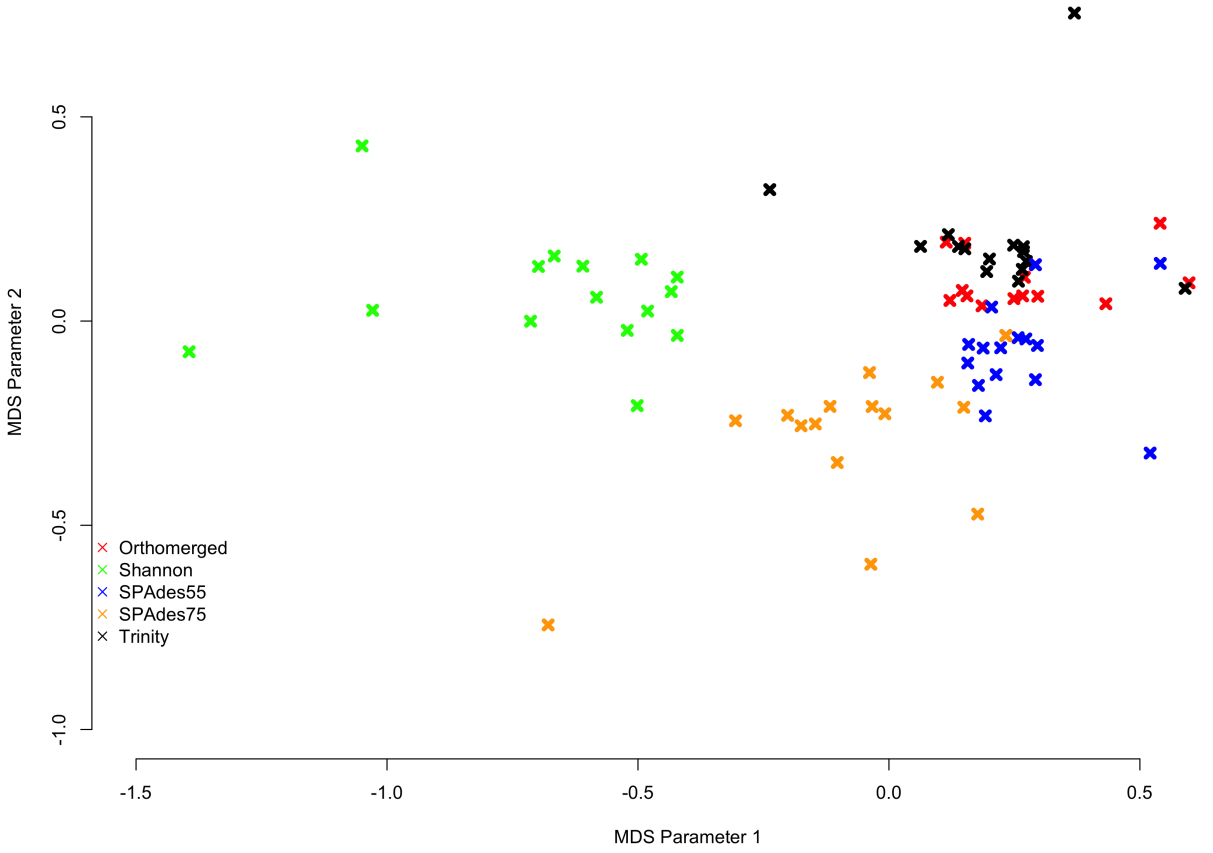

Figure 1. MDS plot describing the similarity within and between assemblers. Colored x's mark individual assemblies, with red marks corresponding to the ORP assemblies, green marks corresponding to the `Shannon` assemblies, blue marks corresponding to the `SPAdes55` assemblies, orange marks corresponding to the `SPAdes75` assemblies, and the black marks corresponding to the `Trinity` assemblies. In general assemblies produced by a given assembler tend to cluster together.

3.1.1 Assembly Structure

The structural integrity of each assembly was evaluated using the `TransRate` and `Detonate` software packages. As many downstream applications depend critically on accurate read mapping, assembly quality is correlated with increased mapping rates. The split violin plot presented in figure 2A visually represents the mapping rates of each assembly, with lines connecting the mapping rates of datasets assembled with `Trinity` and with the ORP, respectively. The average mapping rate of the `Trinity` assembled datasets was

87% (sd = 8%), while the average mapping rates of the ORP assembled datasets was 93% (sd=4%). This test is statistically significant (one-sided Wilcoxon rank sum test, p = 2E-2). Mapping rates of the other assemblies are less than that of the ORP assembly, but in most cases, greater than that of the Trinity assembly. This aspect of assembly quality is critical. Specifically mapping rates measure how representative the assembly is of the reads. If I assume that the vast majority of generated reads come from the biological sample under study, when reads fail to map, that fraction of the biology is lost from all downstream analysis and inference. This study demonstrates that across a wide variety of taxa, assembling RNAseq reads with any single assembler alone may result in a decrease in mapping rate and in turn, the lost ability to draw conclusions from that fraction of the sample.

Figure 2B describes the distribution of `TransRate` assembly scores, which is a synthetic metric taking into account the quality of read mapping and coverage-based statistics. The `Trinity` assemblies had an average optimal score of 0.35 (sd = .14), while the ORP assembled datasets had an average score of 0.46 (sd = .07). This test is statistically significant (one-sided Wilcoxon rank sum test, p-value = 1.8E-2). Optimal scores of the other assemblies are less than that of the ORP assembly, but in most cases, greater than that of the `Trinity` assembly. Figure 2C describes the distribution of `Detonate` scores. The `Trinity` assemblies had an average score of -6.9E9 (sd = 5.2E9), while the ORP assembled datasets had an average score of -5.3E9 (sd = 3.5E9). This test not is statistically significant, though in all cases, relative to all other assemblies, scores of the ORP assemblies are improved (become less negative), indicating that the ORP produced assemblies of higher quality.

In addition to reporting synthetic metrics related to assembly structure, `TransRate` reports individual metrics related to specific elements of assembly quality. One such metric estimates the rate of chimerism, a phenomenon which is known to be problematic in *de novo* assembly (Ungaro, Pech, J.-F. Martin, McCairns, Mévy, Chappaz, and Gilles, 2017; Singhal, 2013). Rates of chimerism are relatively constant between all assemblers, ranging from 10% for the `Shannon` assembly, to 12% for the `SPAdes75` assembly. The chimerism rate for the ORP assemblies averaged 10.5% ($\pm$ 4.7%). While the new method would ideally improve this metric by exclusively selecting non-chimeric transcripts, this does not seem to be the case, and may be related to the inherent shortcomings of short-read transcriptome assembly.

Of note, consistent with all short-read assemblers (Ungaro, Pech, J.-F. Martin, McCairns, Mévy, Chappaz, and Gilles, 2017), the ORP assemblies may not accurately reflect the true isoform complexity. Specifically, because of the way that single representative transcripts are chosen from a cluster of related sequences, some transcriptional complexity may be lost. Consider the cluster containing contigs {AB, A, B} where AB is a false-chimera, selecting a single representative transcript with the best score could yield either A or B, thereby

excluding an important transcript in the final output. I believe this type of transcript loss is not common, based on how contigs are scored (Table 1, Figure 3, ~~R. Smith-Unna, Boursnell, Patro, Hibberd, and Kelly, 2016~~(R. Smith-Unna, Boursnell, Patro, Hibberd, and Kelly, 2016)), though strict demonstration of this is not possible, given the lack of high-quality reference genomes for the majority of the datasets. More generally, mapping rates, `Detonate` and `TransRate` score improvements suggest that this type of loss is not widespread.

Figure 2

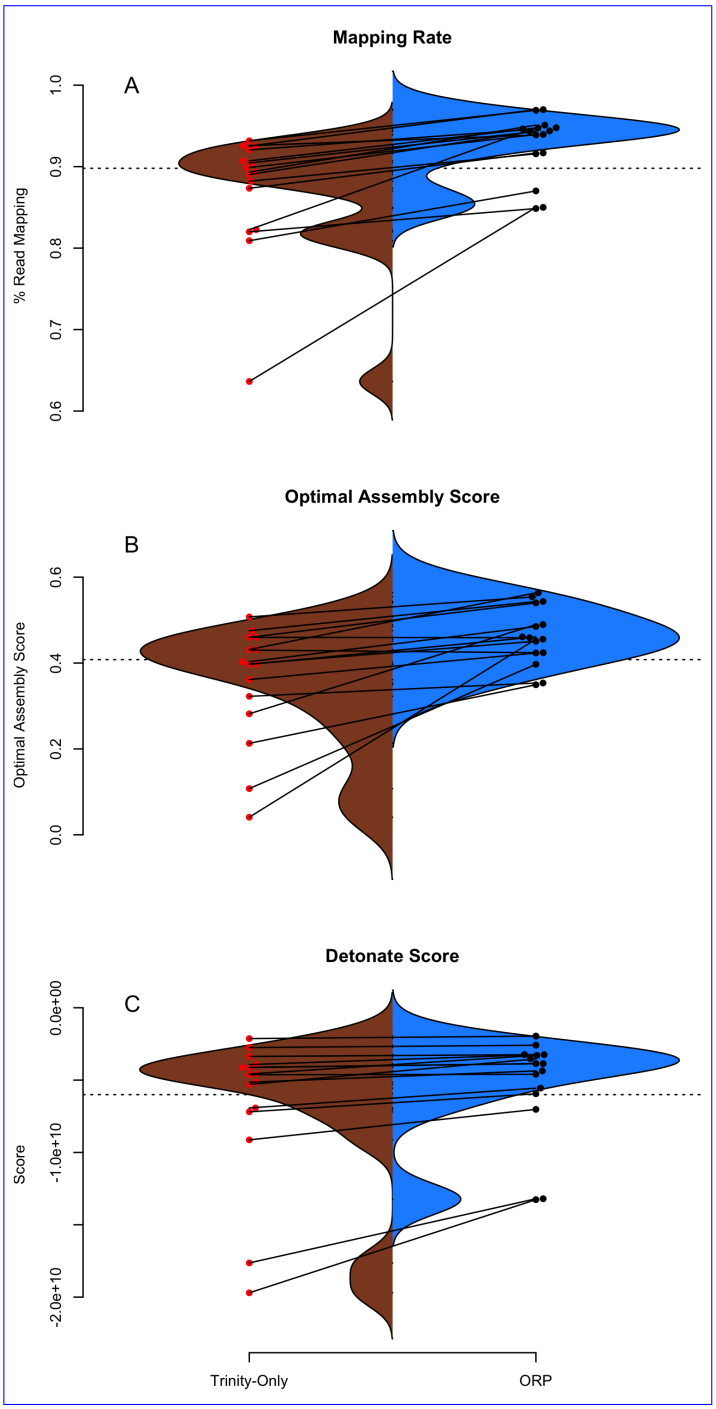

Figure 2. **`TransRate`** and **`Detonate`** generated statistics. Split violin plots depict the relationship between **`Trinity`** assemblies (brown color) and ORP produced assemblies (blue color). Red and black dots indicate the value of a given metric for each assembly. Lines connecting the red and black dots connect datasets assembled via the two methods.

3.1.2 Assembly Content

The genic content of assemblies was measured using the software package `Shmlast`, which implements the conditional reciprocal blast test against the Swiss-prot database. Presented in Table 2 and in Figure 3A, ORP assemblies recovered on average 13364 (sd=3391) blast hits, while all other assemblies recovered fewer (minimum `Shannon`, mean=10299). In every case across all assemblers, the ORP assembler retained more reciprocal blast hits, though only the comparison between the ORP assembly and `Shannon` was significant (one-sided Wilcoxon rank sum test, p = 4E-3). Notably, in all cases, each assembler was both missing transcripts contained in other assemblies, and contributed unique transcripts to the final merged assembly (Table 2), highlighting the utility of using multiple assemblers.

Table 2

Assembly	Genes	Delta	Unique
Concatenated	14674 ± 3590		
SPAdes55		-1739 ± 758	570 ± 266
SPAdes75		-2711 ± 2047	301 ± 195
Shannon		-4375 ± 3508	302 ± 241
Trinity		-1952 ± 803	520 ± 301

Table 2 describes the number of genes contained in the assemblies, with the row labelled concatenated representing the combined average ($\pm$ standard deviation) number of genes contained in all assemblies of a given dataset. The other rows contain information about each assembly. The column labelled delta contains the average number ($\pm$ standard deviation) of genes missing, relative to the concatenated number. The unique column contains the average number of genes ($\pm$ standard deviation) unique to that assembly.

Regarding `BUSCO` scores, `Trinity` assemblies contained on average 86% (sd = 21%) of the full-length orthologs as defined by the `BUSCO` developers, while the ORP assembled datasets contained on average 86% (sd = 13%) of the full length transcripts. Other assemblers contained fewer full-length orthologs. The `Trinity` and ORP assemblies were missing, on average 4.5% (sd = 8.7%) of orthologs. The `Trinity` assembled datasets contained 9.5% (sd = 17%) of fragmented transcripts while the ORP assemblies each contained on average 9.4% (sd = 9%) of fragmented orthologs. The other assemblers in all cases contained more fragmentation. The rate of transcript duplication, depicted in figure 3B is 47% (sd = 20%) for `Trinity` assemblies, and 34% (sd = 15%) for ORP assemblies. This result is statistically significant (One sided Wilcoxon rank sum test, p-value = 0.02). Of note, all other assemblers produce less transcript duplication than does the ORP assembly, but none of these differences arise to the level of statistical significance.

While the majority of the BUSCO metrics were unchanged, the number of orthologs recovered in duplicate (>1 copy), was decreased when using the ORP. This difference is important, given that the relative frequency of transcript duplication may have important implications for downstream abundance estimation, with less duplication potentially resulting in more accurate estimation. Although gene expression quantitation software (Patro, Duggal, Love, Irizarry, and Kingsford, 2017; Bray, Pimentel, Melsted, and Lior Pachter, 2016) probabilistically assigns reads to transcripts in an attempt at mitigating this issue, a primary solution related to decreasing artificial transcript duplication could offer significant advantages.

Figure 3

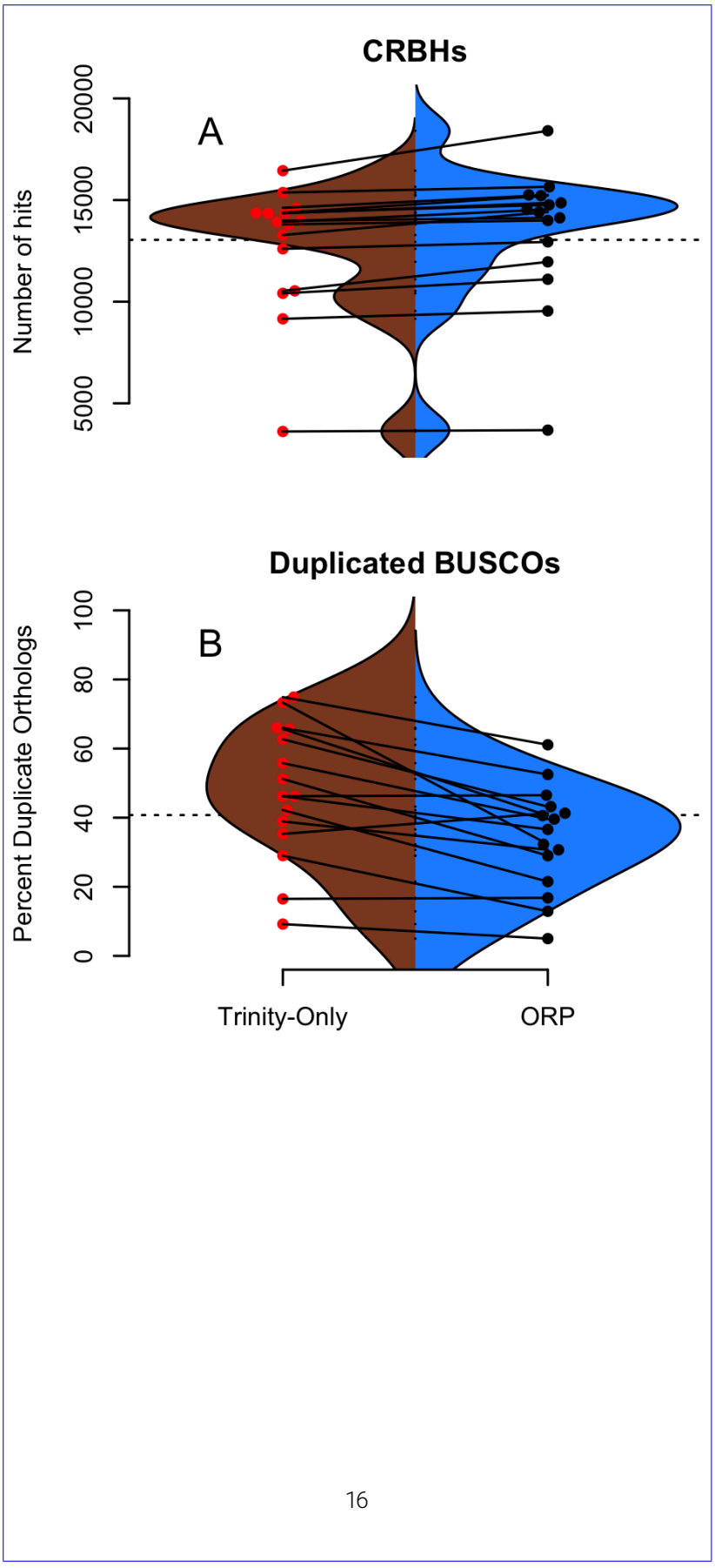
CRBHs
A
Number of hits
5000
10000
15000
20000
Duplicated BUSCOs
B
Percent Duplicate Orthologs
0
20
40
60
80
100
Trinity-Only
ORP

Figure 3. `Shmlast` and `BUSCO` generated statistics. Split violin plots depict the relationship between `Trinity` assemblies (brown color) and ORP produced assemblies (blue color). Red and black dots indicate the value of a given metric for each assembly. Lines connecting the red and black dots connect datasets assembled via the two methods.

3.1.3 Assembler Contributions

To understand the relative contribution of each assembler to the final merged assembly produced by the Oyster River Protocol, I counted the number of transcripts in the final merged assembly that originated from a given assembler (Figure 4). On average, 36% of transcripts in the merged assembly were produced by the `Trinity` assembler. 16% were produced by `Shannon`. `SPAdes` run with a kmer value of length=55 produced 28% of transcripts, while `SPAdes` run with a kmer value of length=75 produced 20% of transcripts

Figure 4

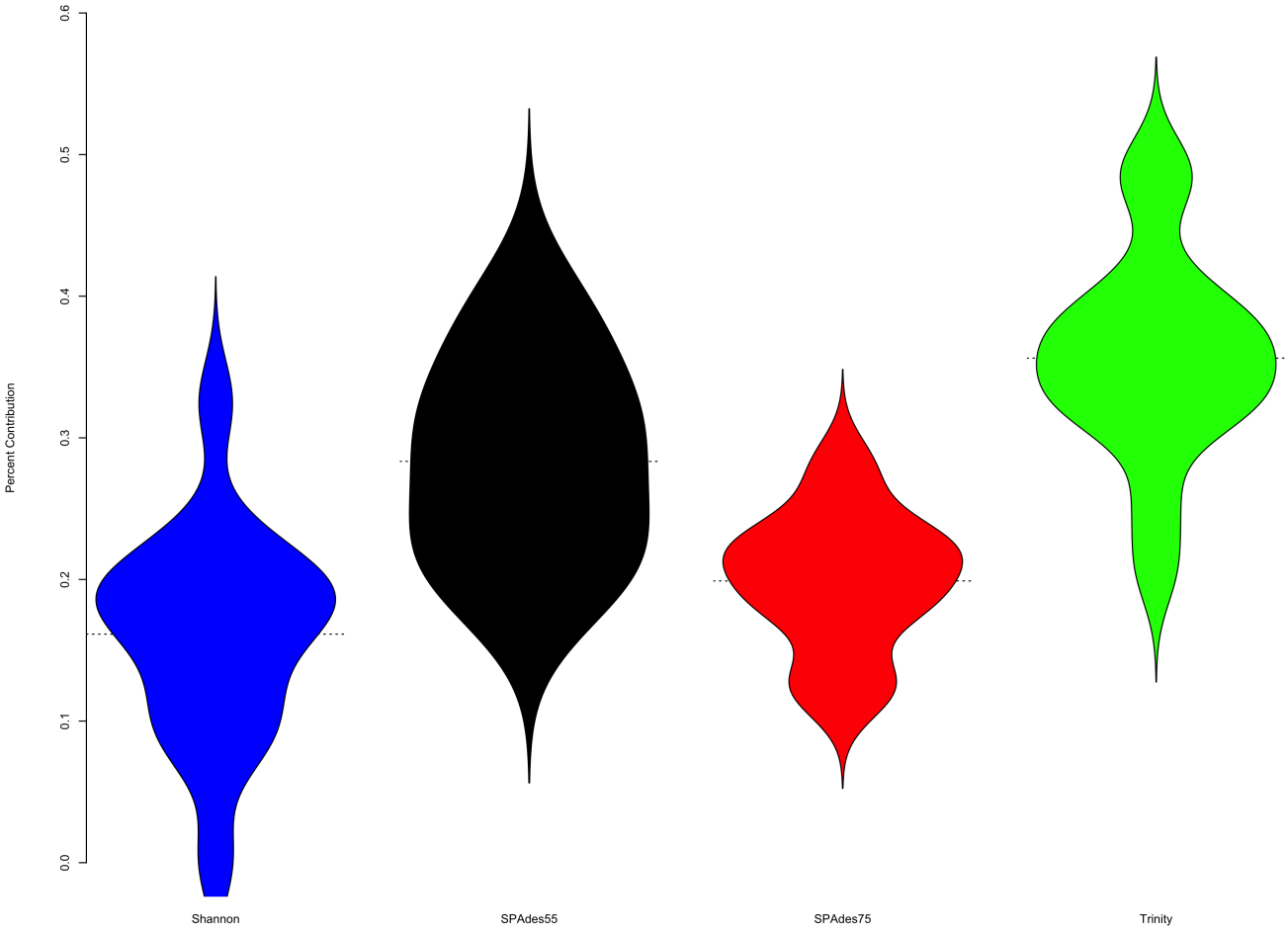

Figure 4 describes the contribution of each assembler to the final ORP assembly. Violin plots illustrate that

Shannon contributes on average the fewest number of transcripts to the final merged assembly, while Trinity contributes on average the most. Small dashed lines on each side of the plot mark the median of the distribution.

To further understand the potential biases intrinsic to each assembler, I plotted the distribution of gene expression estimates for each merged assembly, broken down by the assembler of origin (Figure 5, depicting four randomly selected representative assemblies). As is evident, most transcripts are lowly expressed, with `SPAdes` and `Trinity` both doing a sufficient job in reconstructing these transcripts. Of note, the `SPAdes` assemblies using kmer-length=75 is biased, as expected, towards more highly expressed transcripts relative to kmer-length 55 assemblies. `Shannon` demonstrates a unique profile, consisting of, almost exclusively high-expression transcripts, showing a previously undescribed bias against low-abundance transcripts. These differences may reflect a set of assembler-specific heuristics which translate into differential recovery of distinct fractions of the transcript community. Figure 5 and Table 2 describe the outcomes of these processes in terms of transcript recovery. Taken together, these expression profiles suggest a mechanism by which the ORP outperforms single-assembler assemblies. While there is substantial overlap in transcript recovery, each assembler recovers unique transcripts (Table 2 and Figure 5) based on expression (and potentially other properties), which when merged together into a final assembly, increases the completeness

Figure 5

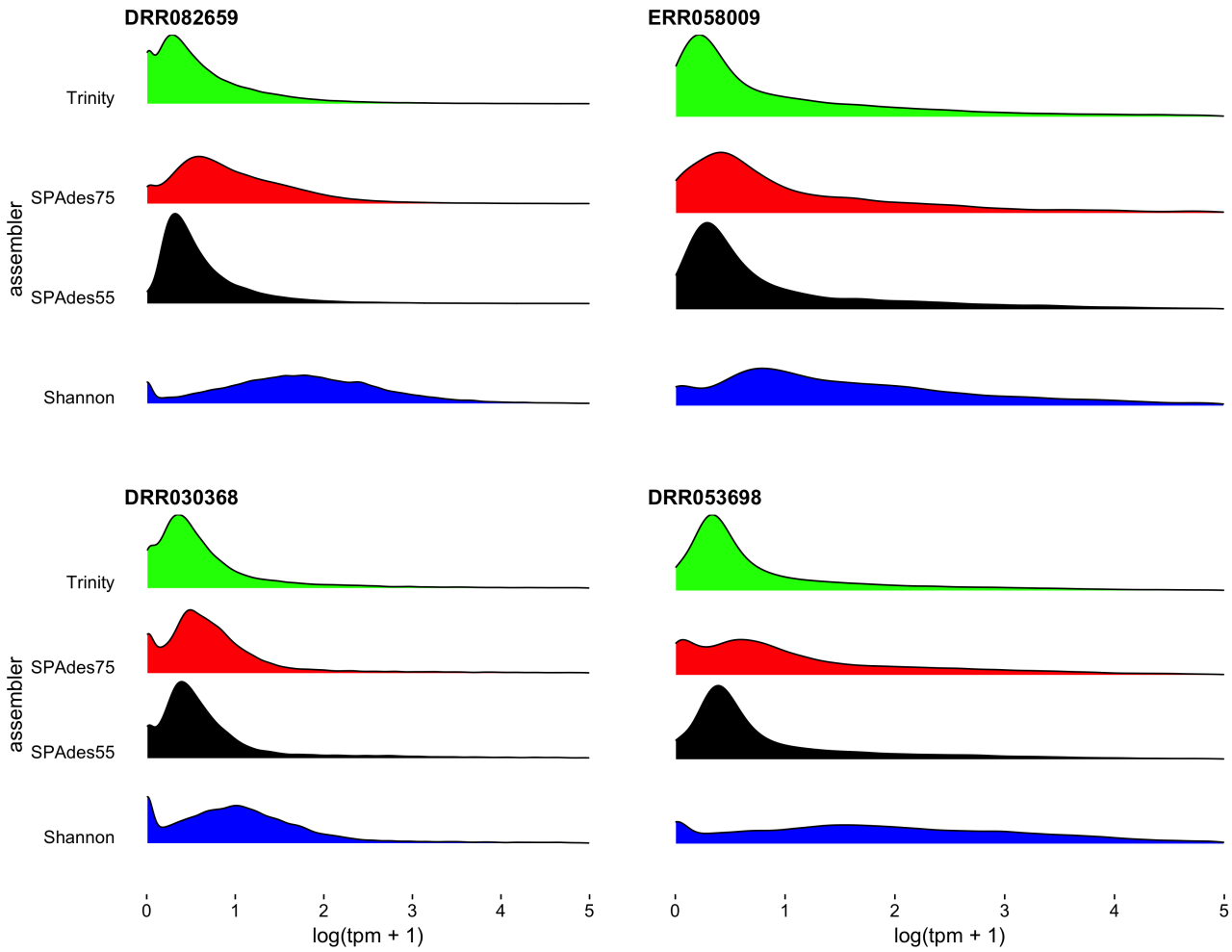

Figure 5 depicts the density distribution of gene expression (log(TPM+1)), broken down by individual assembly, for four representative datasets. As predicted, the use of a higher kmer value with the `SPAdes` assembler resulted in biasing reconstruction towards more highly expressed transcripts. Interestingly, `Shannon` uniquely exhibits a bias towards the reconstruction of high-expression transcripts (or away from low-abundance transcripts).

3.2 Quality is independent of read depth

This study included read datasets of a variety of sizes. Because of this, I was interested in understanding if the number of reads used in assembly was strongly related to the quality of the resultant assembly. Conclusively, this study demonstrates that between 30 million paired-end reads and 200 million paired-end reads, no strong patterns in quality are evident (Figure 6). This finding is in line with previous work, (Matthew D MacManes, 2015) suggesting that assembly metrics plateau at between 20M and 40M read

pairs, with sequencing beyond this level resulting in minimal gain in performance.

Figure 6

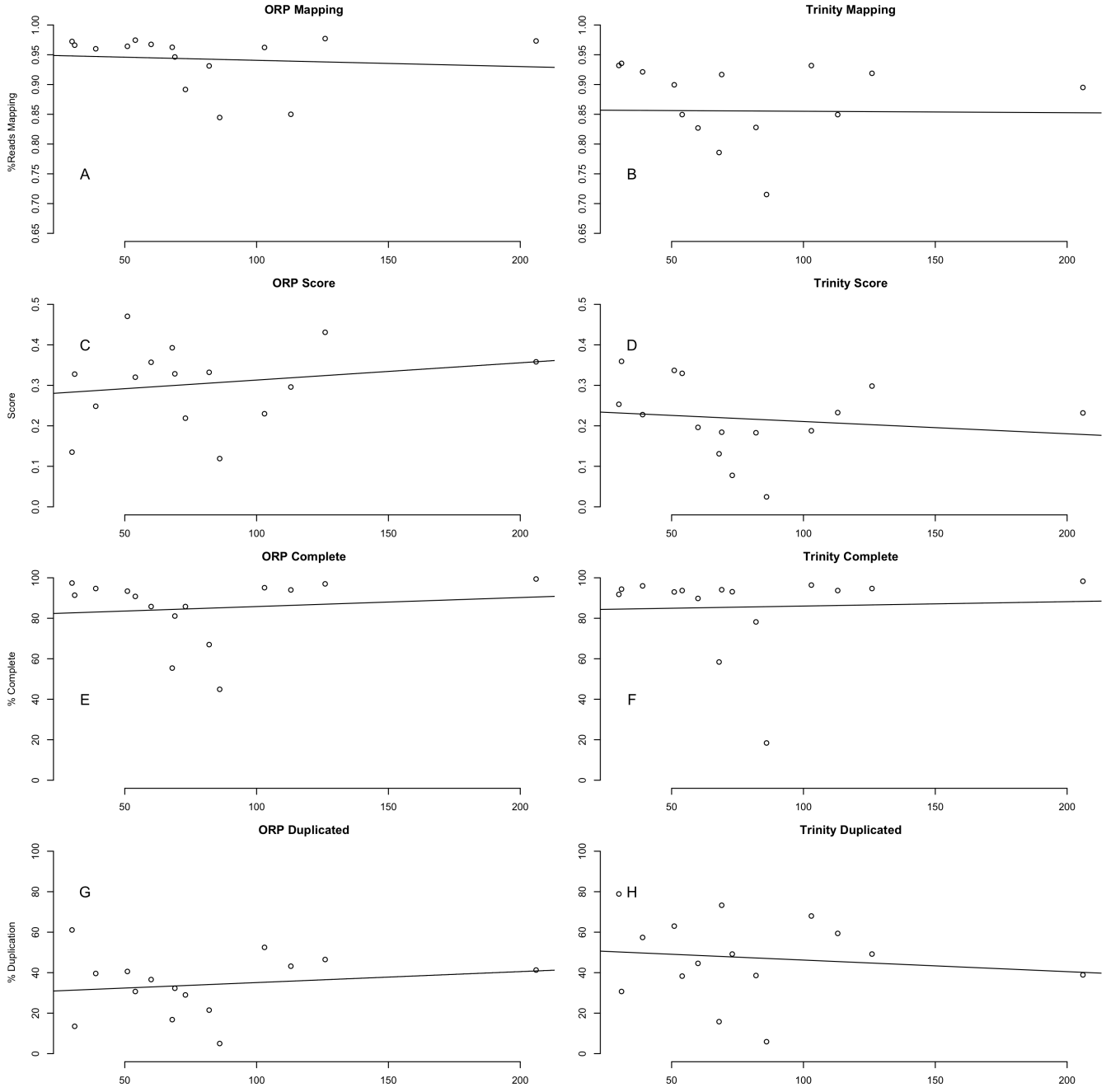

Figure 6 depicts the relationship between a subset of assembly metrics and the number of read pairs. There is no significant relationship. In all cases the x-axis is millions of paired-end reads.

4 Conclusions

~~For non-model organisms lacking reference genomic resources, the error corrected, adapter- and quality-trimmed reads must be assembled *de novo* into transcripts. While the assembly package `Trinity` (Haas et al., 2013) is thought to currently be the most accurate stand-alone assembler (B. Li, Fillmore, Bai, Collins, Thomson, Stewart, and Dewey, 2014), a merged assembly with multiple assemblers results in higher quality assemblies. Specifically, use of the Oyster River Protocol, which contains a recipe for read error correction, quality trimming, assembly with multiple software packages, and merging resulted in a final assembly, the structure of which was greatly improved.~~

~~Specifically, the improvements in assembly metrics described here are attributed to the multi-way approach, where three different assemblers and three different kmer lengths were used. This approach allows the strengths of one assembler to effectively complement the weaknesses of another, thereby resulting in a more complete assembly than otherwise possible. These enhancements are important, as unassembled transcripts are invisible to all downstream analysis.~~

Acknowledgments

This work was significantly improved by discussions with Anthony Westbrook, Richard Smith-Unna ~~, Rob Patro ,~~ and Rob Patro and reviewers C. Titus Brown, Brian Haas~~and many others. More generally, the work and its presentation has been influenced by supporters of the Open Access and Open Science movements~~, and Nick Schurch. MacManes is supported by the NSF (IOS 1455960, OCE 1455960, DEB 1655585), XSEDE (MCB110134), and the UNH Research and Computing Center.

References

Aubry, Sylvain, Steven Kelly, Britta M C Kümpers, Richard D Smith-Unna, and Julian M Hibberd (2014). "Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis". In: *PLOS Genetics* 10.6, e1004365.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, btu170–2120.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter (2016). "Near-optimal probabilistic RNA-seq quantification." In: *Nature Biotechnology* 34.5, pp. 525–527.

Cahoy, John D, Ben Emery, Amit Kaushal, Lynette C Foo, Jennifer L Zamanian, Karen S Christopherson, Yi Xing, Jane L Lubischer, Paul A Krieg, Sergey A Krupenko, Wesley J Thompson, and Ben A Barres (2008). "A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 28.1, pp. 264–278.

Chikhi, Rayan and Paul Medvedev (2014). "Informed and automated k-mer size selection for genome assembly." In: *Bioinformatics* 30.1, pp. 31–37.

Emms, David M and Steven Kelly (2015). "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy." In: *Genome Biology* 16.1, p. 157.

Finseth, Findley R and Richard G Harrison (2014). "A comparison of next generation sequencing technologies for transcriptome assembly and utility for RNA-Seq in a non-model bird." In: *PloS one* 9.10, e108550.

Fitzpatrick, M.J., Y Ben-Shahar, L.E.M. Vet, H.M. Smid, G E Robinson, and M.B. Sokolowski (2005). "Candidate genes for behavioural ecology". In: *Trends In Ecology & Evolution* 20.2, pp. 96–104.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D Leduc, Nir Friedman, and Aviv Regev (2013). "*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." In: *Nature Protocols* 8.8, pp. 1494–1512.

Jackman, Shaun D and Inanç Birol (2016). "Linuxbrew and Homebrew for cross-platform package management [version 1; not peer reviewed]". In: *F1000*.

Jiang, Hongshan, Rong Lei, Shou-Wei Ding, and Shuifang Zhu (2014). "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads." In: *BMC Bioinformatics* 15.1, p. 182.

Kampstra, P (2008). "Beanplot: A boxplot alternative for visual comparison of distributions". In:

Kannan, S, J Hui, K Mazooji, L Pachter, and D Tse (2016). "Shannon: An Information-Optimal de Novo RNA-Seq Assembler". In: *bioRxiv*.

Lappalainen, Tuuli, Michael Sammeth, Marc R Friedländer, Peter A C t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy,

Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigo, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis (2013). "Transcriptome and genome sequencing uncovers functional variation in humans." In: *Nature* 501.7468, pp. 506–511.

Le, H S, M H Schulz, B M McCauley, V F Hinman, and Z Bar-Joseph (2013). "Probabilistic error correction for RNA sequencing". In: *Nucleic Acids Research* 41.10, pp. 1–11.

Li, Bo, Nathanael Fillmore, Yongsheng Bai, Mike Collins, James A Thomson, Ron Stewart, and Colin N Dewey (2014). "Evaluation of *de novo* transcriptome assemblies from RNA-Seq data". In: *Genome Biology* 15.12, pp. 663–21.

Li, R, Y Li, K Kristiansen, and J Wang (2008). "SOAP: short oligonucleotide alignment program". In: *Bioinformatics* 24.5, pp. 713–714.

Li, Xin, Yungil Kim, Emily K Tsang, Joe R Davis, Farhan N Damani, Colby Chiang, Gaelen T Hess, Zachary Zappala, Benjamin J Strober, Alexandra J Scott, Amy Li, andrea Ganna, Michael C Bassik, Jason D Merker, GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Ira M Hall, Alexis Battle, and Stephen B Montgomery (2017). "The impact of rare variation on gene expression across tissues." In: *Nature* 550.7675, pp. 239–243.

Liu, Juntao, Guojun Li, Zheng Chang, Ting Yu, Bingqiang Liu, Rick McMullen, Pengyin Chen, and Xiuzhen Huang (2016). "BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data". In: *PLOS Computational Biology* 12.2, e1004772.

Love, Michael I, Wolfgang Huber, and Simon anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome Biology* 15.12, p. 550.

MacManes, Matthew D (2014). "On the optimal trimming of high-throughput mRNA sequence data." In: *Frontiers in Genetics* 5, p. 13.

MacManes, Matthew D (2015). *Establishing evidenced-based best practice for the de novo assembly and evaluation of transcriptomes from non-model organisms*. Tech. rep.

MacManes, Matthew David and Michael B Eisen (2013). "Improving transcriptome assembly through error correction of high-throughput sequence reads." In: *PeerJ* 1, e113.

Marchant, A, F Mougel, V Mendonça, M Quartier, E Jacquin-Joly, J A da Rosa, E Petit, and M Harry (2016). "Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus." In: *Insect Biochemistry and Molecular Biology* 69, pp. 25–33.

Martin, Marcel (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1, pp. 10–12.

Moreton, Joanna, Abril Izquierdo, and Richard D Emes (2015). "Assembly, Assessment, and Availability of *De novo* Generated Eukaryotic Transcriptomes." In: *Frontiers in Genetics* 6, p. 361.

Mortazavi, Ali, Brian A Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628.

Panhuis, T M (2006). "Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*". In: *Genetics* 173.4, pp. 2039–2047.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford (2017). "Salmon provides fast and bias-aware quantification of transcript expression." In: *Nature Methods* 14.4, pp. 417–419.

Pell, Jason, Arend Hintze, Rosangela Canino-Koning, Adina Howe, James M Tiedje, and C Titus Brown (2012). "Scaling metagenome sequence assembly with probabilistic *de Bruijn* graphs." In: *Proceedings of the National Academy of Sciences* 109.33, pp. 13272–13277.

Peng, Yu, Yu Peng, Henry C M Leung, Henry C M Leung, Siu-Ming Yiu, Ming-Ju Lv, Ming-Ju Lv, Xin-Guang Zhu, Xin-Guang Zhu, Francis Y L Chin, and Francis Y L Chin (2013). "IDBA-tran: a more robust *de novo de Bruijn* graph assembler for transcriptomes with uneven expression levels." In: *Bioinformatics* 29.13, pp. i326–i334.

R Core Development Team, Firstname (2011). "R: A Language and Environment for Statistical Computing". In:

Robertson, Gordon, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S Butterfield, Richard Newsome, Simon K Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna-Liisa Prabhu, Angela Tam, Yongjun Zhao, Richard A Moore, Martin Hirst, Marco A Marra, Steven J M Jones, Pamela A Hoodless, and Inanç Birol (2010). "*De novo* assembly and analysis of RNA-seq data". In: *Nature Methods* 7.11, pp. 909–912.

462 Robinson, Mark D, Davis J. McCarthy, and Gordon K Smyth (2010). “edgeR: a Bioconductor package for
463 differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1, pp. 139–140.
464 Schulz, Marcel H, Daniel R Zerbino, Martin Vingron, and Ewan Birney (2012). “Oases: robust *de novo* RNA-seq
465 assembly across the dynamic range of expression levels.” In: *Bioinformatics* 28.8, pp. 1086–1092.
466 Scott, Camille (2017). “shmlast: An improved implementation of Conditional Reciprocal Best Hits with LAST
467 and Python”. In: *The Journal of Open Source Software* 2.9, pp. 1–4.
468 Simão, Felipe A, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov
469 (2015). “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.”
470 In: *Bioinformatics* 31.19, pp. 3210–3212.
471 Simpson, J T, K Wong, S D Jackman, J E Schein, S J M Jones, and I Birol (2009). “ABySS: A parallel assembler
472 for short read sequence data”. In: *Genome Research* 19.6, pp. 1117–1123.
473 Singhal, Sonal (2013). “*De novo*transcriptomic analyses for non-model organisms: an evaluation of methods
474 across a multi-species data set”. In: *Molecular Ecology Resources* 13.3, n/a–n/a.
475 Smith-Unna, Richard, Chris Boursnell, Rob Patro, Julian M Hibberd, and Steven Kelly (2016). “TransRate:
476 reference-free quality assessment of *de novo* transcriptome assemblies.” In: *Genome Research* 26.8,
477 pp. 1134–1144.
478 Song, Li and Liliana Florea (2015). “Rcorrector: efficient and accurate error correction for Illumina RNA-seq
479 reads.” In: *GigaScience* 4.1, p. 48.
480 Tan, Meng How, Qin Li, Raghuvaran Shanmugam, Robert Piskol, Jennefer Kohler, Amy N Young,
481 Kaiwen Ivy Liu, Rui Zhang, Gokul Ramaswami, Kentaro Ariyoshi, Ankita Gupte, Liam P Keegan,
482 Cyril X George, Avinash Ramu, Ni Huang, Elizabeth A Pollina, Dena S Leeman, Alessandra Rustighi,
483 Y P Sharon Goh, GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis
484 Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups,
485 NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source
486 Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank
487 Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI
488 Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration
489 &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Ajay Chawla,
490 Giannino Del Sal, Gary Peltz, Anne Brunet, Donald F Conrad, Charles E Samuel, Mary A O’Connell,
491 Carl R Walkley, Kazuko Nishikura, and Jin Billy Li (2017). “Dynamic landscape and regulation of RNA
492 editing in mammals.” In: *Nature* 550.7675, pp. 249–254.

Titus Brown, C and Luiz Irber (2016). “sourmash: a library for MinHash sketching of DNA”. In: *The Journal of Open Source Software* 1.5, pp. 27–1.

Ungaro, Arnaud, Nicolas Pech, Jean-François Martin, R J Scott McCairns, Jean-Philippe Mévy, Rémi Chappaz, and andré Gilles (2017). “Challenges and advances for transcriptome assembly in non-model species”. In: *PloS one* 12.9, e0185020–21.

Vijay, Nagarjun, Jelmer W Poelstra, Axel Künstner, and Jochen B W Wolf (2013). “Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments.” In: *Molecular Ecology* 22.3, pp. 620–634.

Wang, Sufang and Michael Gribskov (2017). “Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis.” In: *Bioinformatics* 33.3, pp. 327–333.

Wang, Zhong, Zhong Wang, Mark Gerstein, and Michael Snyder (2009). “RNA-Seq: a revolutionary tool for transcriptomics.” In: *Nature Reviews Genetics* 10.1, pp. 57–63.

Wolf, Jochen B W (2013). “Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial”. In: *Molecular Ecology Resources* 13.4, pp. 559–572.

Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, Xin Zhou, Yingrui Li, Xun Xu, Gane Ka-Shu Wong, and Jun Wang (2014). “SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads.” In: *Bioinformatics* 30.12, pp. 1660–1666.

Yang, Xiao, Xiao Yang, Karin S Dorman, Karin S Dorman, Srinivas Aluru, and Srinivas Aluru (2010). “Reptile: representative tiling for short read error correction.” In: *Bioinformatics* 26.20, pp. 2526–2533.

Yang, Ya and Stephen A Smith (2013). “Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics.” In: *BMC Genomics* 14, p. 328.

Zerbino, D R and E Birney (2008). “Velvet: Algorithms for *de novo* short read assembly using *de Bruijn* graphs”. In: *Genome Research* 18.5, pp. 821–829.