Comparative Biochemistry and Physiology, Part C 155 (2012) 95–101

Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology, Part C

journal homepage: www.elsevier.com/locate/cbpc

Effects of short read quality and quantity on a de novo vertebrate transcriptome assembly ☆

T.I. Garcia [a], Y. Shen [a], J. Catchen [b], A. Amores [b], M. Schartl [c], J. Postlethwait [b], R.B. Walter [a,*]

[a] Department of Chemistry and Biochemistry, 419 Centennial Hall, Texas State University, 601 University Drive, San Marcos, TX 78666, USA
[b] Institute of Neuroscience, University of Oregon, 1425 E. 13th Avenue, Eugene, OR 97403, USA
[c] Universität Würzburg, Physiologische Chemie I, Biozentrum, Am Hubland, D-97074 Würzburg, Germany

ARTICLE INFO

Article history:
Received 10 February 2011
Received in revised form 20 May 2011
Accepted 24 May 2011
Available online 1 June 2011

Keywords:
NGS
Short read
Quality
Phred
Assembly
Quantity
Velvet

ABSTRACT

For many researchers, next generation sequencing data holds the key to answering a category of questions previously unassailable. One of the important and challenging steps in achieving these goals is accurately assembling the massive quantity of short sequencing reads into full nucleic acid sequences. For research groups working with non-model or wild systems, short read assembly can pose a significant challenge due to the lack of pre-existing EST or genome reference libraries. While many publications describe the overall process of sequencing and assembly, few address the topic of how many and what types of reads are best for assembly. The goal of this project was use real world data to explore the effects of read quantity and short read quality scores on the resulting de novo assemblies. Using several samples of short reads of various sizes and qualities we produced many assemblies in an automated manner. We observe how the properties of read length, read quality, and read quantity affect the resulting assemblies and provide some general recommendations based on our real-world data set.

Published by Elsevier Inc.

1. Introduction

Current next-generation sequencing (NGS) technologies enable researchers to address myriad questions regarding biological and genetic mechanisms. NGS enables researchers to rapidly sequence genomes (Bentley et al., 2008; Li et al., 2010) or transcriptomes, to obtain snapshots of global gene expression levels in RNA-seq experiments (Wang et al., 2009; Costa et al., 2010) and to examine genome-wide protein-DNA interactions in ChIP-seq experiments (Barski and Zhao, 2009; Park, 2009) among others. Advancing technologies and economies of scale have collaborated to bring these capabilities within reach of even small research communities studying non-model or wild systems and organisms. Even so, there exist many barriers to entry into such studies that can make it difficult to initiate NGS projects or to extract meaning from the data. In this project we address some of the questions researchers may have when embarking on a new NGS project with regard to both quantity and quality of short reads needed for assembly.

There are many NGS platforms available (Metzker, 2009; Voelkerding et al., 2009; Bräutigam and Gowik, 2010; Nowrousian, 2010), but the most common and prolific NGS methods available today produce extremely large quantities of very short reads (Illumina or ABI SOLiD sequencing platforms). Successfully assembling these short reads into a set of contigs representing the original sequences in the biological sample is a complex problem. This problem is easiest to solve if a reference genome or transcriptome is available to guide the assembly. However, in the absence of a reference genome, one must resort to de novo assembly of short-read data, which is more difficult. More challenging still is de novo assembly of data derived from source material that is expected to have uneven coverage, such as RNA transcripts where message abundance varies by several orders of magnitude, and there may be multiple versions of each message. For many researchers working with non-model organism transcriptomes, the last situation is the norm.

Many of these technologies are capable of producing single or paired-end reads. Paired-end read information can help to resolve repetitive regions which might otherwise be intractable to the assembly program (Narzisi and Mishra, 2011; Wetzel et al., 2011). They also allow scaffolding of contigs which would otherwise remain fragmented and they aid in the identification of splice junctions and alternative splice forms. They can also be beneficial to further analyses downstream of assembly.

Of the variety of NGS technologies available, we will focus here on RNA transcript paired-end short reads collected from the Illumina Genome Analyzer platform and assembled de novo without the aid of

☆ This paper is based on a presentation given at the *5th Aquatic Annual Models of Human Disease* conference: hosted by Oregon State University and Texas State University-San Marcos, and convened at Corvallis, OR, USA September 20–22, 2010.

* Corresponding author. Tel.: +1 512 245 0358; fax: +1 512 245 1922.
E-mail address: RW12@txstate.edu (R.B. Walter).

1532-0456/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.cbpc.2011.05.012

a reference genome or reference transcriptome. The raw sequence reads produced by this platform are commonly in the range of 60–150 bp in length. The inherent errors in this sequencing technique are generally well-understood (Dohm et al., 2008), and there exist many open-source and commercial packages designed to assemble these data into contigs (Metzker, 2009; Miller et al., 2010; Nowrousian, 2010). A discussion of the many available assembly algorithms is beyond the scope of this paper and has been covered before (Pop, 2009; Paszkiewicz and Studholme, 2010). Here, we have chosen to use the Velvet short-read assembler (Zerbino and Birney, 2008; Zerbino et al., 2009; Zerbino, 2010) because it is widely used, and its abilities and limitations are relatively well understood. We will also use the Oases extension for Velvet which is designed to assemble transcript data specifically and is a much less well documented software package. While Oases is still a relatively 'young' program, it addresses problems inherent in packages tuned to assemble genome data. A recent publication critically assessed its performance with respect to other assembly packages and found it to be the best tool to assemble the chickpea transcriptome (Garg et al., 2011).

Over the course of one year, we produced several sets of transcript data using the Illumina Genome Analyzer platform. Over this time period we observed an increase in both the lengths of reads that could be reliably sequenced as well as the overall average quality scores reported by the sequencer software. While there were several changes to the software, firmware, and sequencing chemistry made available by Illumina, clearly factors outside the sequencing process may also have contributed to the observed differences in read quality. This real-world data set presents a variety of relative qualities and characteristics similar to data encountered by researchers both new and veteran to this field. As such it is difficult to provide strict controls for each variable within the data we have. Nevertheless we are able to make several useful observations that may serve to guide decisions regarding the amount of data needed to obtain an assembly of reasonable quality with a minimal investment.

2. Methods

2.1. Transcript sequencing

Transcript RNA was prepared from 13 separate tissues/organs and life stages of the live-bearing fish, *Xiphophorus maculatus* Jp 163 A. The *X. maculatus* Jp 163 A line is maintained at the *Xiphophorus* Genetic Stock Center by brother–sister matings and is inbred over 100 generations (Walter et al., 2006) (see http://xiphophorus.txstate.edu). RNA samples were sequenced using the Illumina Genome Analyzer platform; the 13 samples were sequenced in 3 batches, submitted at 3 separate time points over a year (Catchen, J., et al., unpublished results). Each tissue or life stage was sequenced in an individual flow cell lane as paired end reads of 36 bp or 60 bp. Overall average quality scores were calculated for each sample, ignoring any quality score values of 2 (encoded as a 'B' in FASTQ format), which is used as a special flag to indicate that some type of sequencing error had occurred. The percentage of reads in a sample containing the 'B' flag was also reported. The short reads were grouped by their time of creation into three supersets designated as L1, L2, and L3. The tissues/stages contained in each sample is as follows: L1 contains whole body RNA from 5 age and/or gender specific samples from 5 days to 15 months of age, L2 contains whole body RNA from two embryonic stages and two tissue samples, L3 contains three tissue samples.

2.2. Quality filtration

We processed raw transcript sequence data to remove low-quality reads by developing a four-stage filtration algorithm for the FASTQ-encoded data set that is comprised of several stages of checks and modifications. Stage 1 is a check for uncalled bases; the default setting is to reject any reads with more than two uncalled bases, but this can be altered at run-time. Velvet converts uncalled bases ('N' characters) into 'A' characters (adenine bases), but in most cases, this would be no different than a random read error. So, instead of rejecting reads with 2 or fewer uncalled bases, we opted to use the default setting to err for increased read coverage. Stage 2 searches for the presence of 'B' characters in the quality scores; these indicate that a particular error event is likely to have occurred and therefore the remainder of the read (distal to the 'B' character) should not be used. Any positions after and including the first 'B' character are trimmed off the sequence. Stage 3 scans for low-quality regions, deleting them and leaving high-quality fragments. Stage 3 first examines each position with a quality equal to or below 20 (1 error in 100). If the mean of the quality scores of the position in question and its up and downstream neighbors is still 20 or below, that position is marked for deletion. Any position with a quality score of 10 (1 error in 10) or below is marked for deletion without the possibility of rescue by neighboring positions. All the marked positions are then deleted, possibly breaking the read into smaller fragments of high quality. In Stage 4, the largest of the fragments remaining from the original read is selected as the trimmed read and the rest is discarded.

In addition to the process of filtering reads, the tool set we developed also measures statistics on the final outcomes of each read. Reads are sorted into four main categories: 1) failure due to quantity of uncalled bases, 2) failure due to size restrictions post trimming, 3) passage with trimming, and 4) passage without trimming. The total number of reads in each of those categories is tallied post filtration. Additionally those reads that passed may have lost or retained their mate in the filtration process and this is noted. The failure rate of a set of reads indicates the fraction of reads that failed for any reason, and the average failure rates calculated for Fig. 1B are the average of failure rates of each component in the L1, L2, and L3 samples.

2.3. Assembly and analysis

We chose to use the Velvet short-read assembler because it is widely used, well-documented, and expertise in its use exists in many institutes. Many parameters can be fine-tuned to improve aspects of a de novo assembly, but it is difficult to generalize their use into an algorithm that gives the best assembly for all data sets. In the manual refinement of an assembly, intuition and trial and error can be important components of the process. However, in order to avoid introducing user-bias into the assembly process, we used a script included with the Velvet package called VelvetOptimiser that optimizes two of the most important settings based on predefined rules.

Version 1.0.14 of Velvet (Zerbino and Birney, 2008) was installed on a Dell R910 rack-mount server with 1 TB of physical memory and 8 quad-core Xeon processors to carry out assemblies. All samples were assembled independently using VelvetOptimiser version 2.1.7 set to select the k-mer size in the range of 21 to 55 bp (a range of 21 to 35 bp was used for the sample with 36 bp reads) that produced the best N50 (contig length-weighted median) value; and then to select a coverage-cutoff that maximized the number of base pairs in large contigs. A further assembly step was carried out by submitting the same read set and kmer size determined by VelvetOptimiser to the Oases (version 0.1.18) extension to Velvet (REF) with an insert length parameter of 200 bp given.

Determining the quality of an assembly can become very involved after several rounds of refinement, but initial efforts can be guided by some basic metrics describing the general size characteristics of the assembled sequences. A target number of assembled transcripts is very difficult to provide a priori. While there is no published *Xiphophorus maculatus* genome, *Oryzias latipes* is a closely related species with a well annotated genome that contains approximately 20,000 genes. However, even with this number as an estimate, alternative splice forms and