Transcriptome Assembly and Evaluation, using Sequencing Quality Control (SEQC) Data

Introduction

The US Food and Drug Administration (FDA) has coordinated the Sequencing Quality Control project (SEQC/MAQC-III) with the goal of assessing the technical performance of RNA-Seq experiments comprehensively. The SEQC consortium has generated benchmark datasets of well-studied reference samples, sequenced at multiple sites, and using different sequencing platforms, with controlled settings. The generated RNA-Seq data is used in separate studies to measure quality metrics, spike-in controls, limits of detection, effects of analytic pipeline and assessments of RNA-Seq accuracy and reproducibility. All samples were distributed among six independent centers in the study: 1- Australian Genome Research Facility (AGR), 2- Beijing Genomics Institute (BGI), 3- Weill Cornell Medical College (CNL), 4- City of Hope (COH), 5- Mayo Clinic (MAY) and 6- Novartis (NVS). The SEQC uses the samples from MAQC I consortium (Shi et. al., 2006): sample A is the well-characterized Universal Human Reference RNA (UHRR), and B is Human Brain Reference RNA (HBRR). The synthetic RNA from the External RNA Control Consortium (ERCC) (Baker et. al, 2005) was spiked in. Samples C and D were generated by mixing samples A and B in ratios of 3:1 and 1:3, respectively. Each one of samples A and B had 5 replicates. Replicates 1 to 4 were prepared in each site. The vendor prepared the fifth replicate. To examine the effect of the instrument on the RNA-Seq experiments, all the samples were sequenced using Illumina's HiSeq 2000, and for generating longer reads three sites sequenced samples A and B using the Roche 454 GS FLX platform. The other next generation sequencing technology, SOLiD, was also used. The SEQC consortium overall sequenced 108 libraries on a HiSeq 2000, 68 libraries on SOLiD, and 6 libraries on a Roche 454, generating more than 100 billion reads, for samples A to D.

The first study that focused on RNA-Seq assessments was recently published (Su et. al. 2014). This study systematically examined the impact of site-specific bias in detecting differentially expressed genes, using different RNA-Seq analysis methods. They showed that none of the tested technologies provided reliable absolute quantification and relative expression measures that agreed well across validation platforms: RNA-seq, qPCR and microarrays. They observed sensitivity of results to analysis pipeline choice, and provided a suggestion for sequencing depth: "An effective sequencing depth is clearly contingent on the experimental goals, with simple gene-level expression profiling only requiring 5–50 million single-ended reads for an appropriate analysis pipeline" (Su et. al.). The RNA-Seq mapping and differential expression testing pipeline such as TopHat2 and CuffDiff (Kim et. al, 2013), Magic (Thierry-Mieg et. al, 2006), BitSeq (Glaus et. al, 2012), Subread (Liao et. al., 2013) and r-make incorporating STAR (Dobin et. al, 2013) were examined and their performance was compared. They also pointed that the SEQC reference datasets are invaluable for a systematic characterization of measurements, and for making reliable conclusions from large-scale experiments.

Proposed Assembly Study, using SEQC Datasets

Motivation

The availability of SEQC datasets provides an excellent opportunity for the scientific community to examine RNA-Seq experiments, and to learn more about its variety of characterizations and applications. The SEQC consortium has focused on RNA-Seq analysis pipelines for differential expression detection, splice junction discovery, sample differences, etc. However, the performance of different SEQC RNA-Seq datasets has not been evaluated for transcriptome assembly (to the best of our knowledge). We believe it is crucial to understand the performance of transcriptome assemblies to improve current practices. Understanding the factors that affect transcriptome assembly is also very important. To achieve these goals, reliable input data, which were generated with the highest standards, are necessary. We propose using SEQC RNA-Seq data for assembling the human transcriptome. The comparison of our resulting *de novo* assembled transcriptomes with the well-annotated human transcriptome can provide insights to the pros and cons of particular procedures used for producing and analyzing SEQC data.

Study Design

The SEQC data was generated at multiple sites and included well-studied samples. The samples A and B differ since sample A (UHRR) consists of ten pooled cancer cell lines, and sample B (HBRR) is from multiple brain regions of 23 donors. Our objective is to examine the effects of different sites, samples and sequencing depths on the assembled transcriptome. We will use the samples sequenced by the Illumina platform. We selected samples A and B (sample-type effect), across all six sequencing sites (site effect), and assemblies will be done for different number of pooled replicates (sequencing depth effect). There are four replicates available for samples A and B, sequenced in each center. The fifth replicate is prepared by the vendor and sequenced in each site. We will use replicates for studying the effects of coverage on the assembly: using only one replicate for the minimum sequencing depth, pooling two (and then three) replicates to increase the coverage, and finally pooling four replicates to gain the maximum coverage for each sample in each site. The vendor-prepared fifth replicate will be used for the assembly independently, due to its different preparation (library prep effect). The examination of all transcriptome assemblies will provide insights (and possibly recommendations) on successful RNA-Seq experiments and transcriptome assemblies. After each assembly, we have designed a multi-step procedure for checking the quality of each output. The proposed assemblies using different samples and coverages are:

- Sample A
    - Assembly 1: Sample A, replicate 1 (minimum sequencing coverage, within each site)
    - Assembly 2: Sample A, replicates 1,2 (2 reps are pooled to increase the coverage)
    - Assembly 3: Sample A, replicates 1,2,3 (3 reps are pooled)
    - Assembly 4: Sample A, replicates 1,2,3,4 (4 reps are pooled, maximum sequencing coverage, per sample within each site)
    - Assembly 5: Sample A, replicate 5 (prepared by Illumina, different from replicates 1 to 4)

- Sample B
    - Assembly 6: Sample B, replicate 1 (minimum sequencing coverage, within each site)
    - Assembly 7: Sample B, replicates 1,2 (2 reps are pooled to increase the coverage)
    - Assembly 8: Sample B, replicates 1,2,3 (3 reps are pooled)
    - Assembly 9: Sample B, replicates 1,2,3,4 (4 reps are pooled, maximum sequencing coverage, per sample within each site)
    - Assembly 10: Sample B, replicate 5 (prepared by Illumina, different from

replicates 1 to 4)

These 10 assemblies will be done for data sets from each of the 6 sites that generated SEQC data. Thus, 60 assemblies are needed to include samples A and B, using different replicate combinations (depths), from each of the 6 sites.

Quality Assessments and Comparisons

The SEQC transcriptome assemblies will be examined to understand the effects of multi-site, pooled-replicate and variable-sample on the transcriptome assembly. The process includes multiple quality assessment steps. Common assembly assessment statistics, such as maximum contig length and N50 will be used to determine the quality of the results, and to compare the assemblies using different parameters. However, as other studies noted (Parra et. al, 2007) (Li et. al., 2014), using additional measures for quality control and comparison of assemblies is important. Relying only on the assembly statistics can be misleading and result on overstating the power of the assemblers. Therefore, we will include the following steps in our validation/assembly comparison pipeline:

- Examining the statistics of each assembly and ranking assemblies accordingly.
- Mapping contigs to the human reference genome as a method to find similarities between our results and the reference.
- Comparing the completeness of assemblies, using Core Eukaryotic Genes Mapping (CEGMA) pipeline (Parra et. al, 2007)
- Assigning scores to the assemblies using DETONATE (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation) (Li et. al, 2014)

Employing more quality control (QC) metrics and comparison/ranking tools ensures the quality of the outputs, and will assist us to find the differences between assemblies based on samples A and B sequenced in different sites with different coverage.

Samples

The SEQC datasets were made publicly available, after publishing the corresponding papers. We have downloaded the datasets for Samples A and B and transferred part of it to PSC for initial tests. All the replicates for these samples are obtained and will be used for transcriptome assemblies.

Computational Approach

In our initial quality control steps, any reads that had adapter sequences attached to them will be discarded. For adapter trimming the Cutadapt software (Martin 2012) will be used. The adapter-free reads will be used for transcriptome assembly. Transcriptome assembly of the SEQC data will be done using Trinity (Haas et. al, 2013), which is a state-of-the-art *de novo* transcriptome assembly pipeline. Trinity uses a de novo algorithm developed specifically for reconstructing the transcriptome using de Bruijn graphs. Transcriptome assembly is challenging mainly because RNA-Seq data coverage levels are not evenly distributed. Furthermore, alternative splicing complicates assembly from individual genes. The goal of the Trinity package is to deliver one graph per expressed gene. Trinity consists of three parts: 1) Inchworm, 2) Chrysalis, and 3) Butterfly. During these three steps, Trinity makes linear contigs from RNA-Seq reads, generates and expands de Bruijn graphs, and finally outputs the transcripts and isoforms. The process starts by decomposing the reads into small overlapping pieces called k-mers and extending them by coverage. Finding the common sections of the intermediate transcripts determines alternative splicing, and those transcripts are re-grouped. The de Bruijn graphs are generated by integration of isoforms that are similar except for one base. Finally, by finding and expanding the common section of transcripts and representing the most compact path on the graph, Trinity delivers the fully assembled transcriptome. The outputs are saved in a FASTA format, which includes all the transcripts.

In the quality control step of the project, each assembly goes through a comprehensive pipeline of quality measurement tools. Each software provides the statistical information for its outputs, and we will use them in our QC and ranking of the outputs. The statistics of assemblies will be used to assign ranking to the result, e.g. between similar assemblies, the one with larger N50 will receive higher ranking. Additionally, mapping results using GMAP (Wu and Watanabe, 2005) or NUCmer (Delcher et. al, 2002), and employing a SNP detection process are important for ranking the assemblies. One of important goal is to adapt the QUAST pipeline (Gurevich et. al, 2013) for evaluating transcriptome assembly. QUAST incorporates Assess Assembly and GAGE, both used for genome assembly validation purposes. We also plan to use CEGMA and DETONATE in our QC and comparisons.

The considerable computational needs of the proposed projects, in particular the large shared memory required for these assemblies, are beyond the computing capabilities at our institution. Therefore, we request SUs on the large shared memory (LSM) and regular shared memory (RSM) nodes on the Bridges system at PSC. We also request Extended Collaborative Support Service (ECSS) for developing a formal workflow for the assembly and QC pipeline. We believe automating the procedure will save time running the software pipelines, reduce the possibility of user-induced errors, and facilitate pipeline optimization on Bridges. We would be welcome the opportunity to work with the XSEDE support team to identify potential improvements in our procedures.

Resource Justification

Trinity is by far the most computationally demanding tool in our workflow. We have run six test assemblies on Greenfield, using one, two and three replicates from Cornell (CNL), for samples A and B. These test runs are used for determining the amount of memory, wall time, and storage space that will be needed for SEQC project. To estimate our resource requirements, we used the maximum number of days that it took to complete samples A and B, using different replicate numbers. For the runs with 4 replicates, we estimated the total number of days, based on the pattern from the 1 to 3 replicate runs. All of the test assemblies fit within the memory of a single 768 GB socket (15 cores) on Greenfield. We anticipate that all jobs will fit within 768 GB on Bridges as well. Given that we must run dozens of these assemblies, the many 3 TB large shared memory (LSM) nodes on Bridges are ideal for our Trinity assemblies. The QC steps do not require large memory, and can utilize the regular shared memory (RSM) nodes on Bridges. RSM nodes are not available on Greenfield for testing, but based on past experience, we expect each QC workflow for a given assembly to finish within 24 hours on a 28-core RSM node.

Bridges Service Units:

Our test runs used 1, 2 or 3 replicates to estimate resource requirements for assembly with increasing sequencing depth. We used both samples A and B in our tests. The longest run took around 4.5 days to complete using 1 socket (768 GB) on Greenfield. We expect the 4-replicate Trinity assemblies will take ~6.5 days to finish, and will be able to fit within a single 768 GB Bridges socket. Tables I-III represent our SU estimations.

TABLE I. SUs Required for SEQC Assemblies (LSM Nodes)

Trinity	***Hours***	***GB/ Assembly***	***# of Assemblies***	***GB Hours***	***SUs (1 SU = 51.2 GB-hours)***
1 replicate	1.5*24	768	24	663,552	12,960
2 replicates	3.5*24	768	12	774,144	15,120
3 replicates	4.5*24	768	12	995,328	19,440
4 replicates	6.5*24	768	12	1,437,696	28,080
		Total	**60**	**3,870,720**	**75,600**

TABLE II. SUs Required for SEQC Validations and Quality Control Completion (RSM Nodes)

QC: CEGMA, QUAST, DETONATE	***Hours/QC workflow***	***Cores***	***# of QC workflows***	***SUs***
All assembled data	24	28	60	40,320

TABLE III. Total Requested SUs for SEQC Assembly and Validations Project

Nodes	**Total SUs**
Large Shared Memory (LSM)	**75,600**
Regular Shared Memory (RSM)	**40,320**

Storage Space on Pylon:

We anticipate that the output from Trinity will be about equal to the raw input data (after deleting unneeded files). The average input and output data size (averaging across sizes for different numbers of replicates) will be ~1 TB. It will be important to keep and analyze the input and output data from the initial assembly throughout the QC and validations steps. Table IV summarizes our storage request for Pylon.

TABLE IV. Storage Space Needed for SEQC Assembly and Validations Project

Algorithm	***Average Input data***	***Average assembly output data***	***# of data sets***	***Storage Spaced***
Trinity Transcriptome Assembly	1 TB	1 TB	60	120 TB
QC: CEGMA, QUAST, DETONATE	Use inputs and outputs from assembly stage	0.5 TB	60	30 TB
			Total Storage on Pylon	**150 TB**

Local Computing Environment: We have a 100 node, 24-core per node, Linux cluster available to us at our institution. None of the available nodes have sufficient shared memory to run the Trinity assemblies we propose to do in this work. Access to the many LSM nodes and the Pylon storage resource on Bridges will be essential to completing this project.

Other Supercomputing Resources: We do not have access to allocations on any other supercomputing resources.

References

Shi, L. et al., The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology. 24, 1151–1161 (2006).

Baker, S.C. et al., The External RNA Controls Consortium: a progress report. Nature Methods 2, 731–734 (2005).

Su Z., Labaj P.P., Li S., Thierry-Mieg J., Thierry-Mieg D., Shi W., Wang C., Schroth G.P., Setterquist R.A., Thompson J.F., et al. (SEQC/MAQC-III Consortium), A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology, 32(9), 903-914 (2014).

Kim, D. et al., TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 14, R36 (2013).

Glaus, P., Honkela, A. and Rattray, M., Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 28, 1721–1728 (2012).

Thierry-Mieg, D. and Thierry-Mieg, J., AceView: a comprehensive cDNA-supported gene and transcripts. Genome Biology. 7, S12 (2006).

Liao, Y., Smyth, G.K. & Shi, W.,The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41, e108 (2013).

Dobin, A. et al.,STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

Parra G., Bradnam K. and Korf I., “CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes”, Bioinformatics, vol. 23, no. 9, pp. 1061-1067 (2007).

Li B., et. al, Evaluation of de novo transcriptome assemblies from RNA-Seq data, bioRXiv, doi: http://dx.doi.org/10.1101/006338 (2014).

Haas, B. J. et al., De novo transcript sequence reconstruction from RNAseq using the Trinity platform for reference generation and analysis. Nature Protoc 8, 1494-1512, doi:10.1038/nprot.2013.084 (2013).

Martin M., Cutadapt removes adapter sequences from high-throughput sequencing reads, Bioinformatics in Action, 17(1), pages 10-12 (2012)

Wu T. D., and Watanabe C. K., GMAP: a genomic mapping and alignment program for mRNA and EST sequences, Bioinformatics, 21, 1859-1875, (2005)

Delcher A.L., Phillippy A., Carlton J., and Salzberg S. L., Fast Algorithms for Large-scale Genome Alignment and Comparision, *Nucleic Acids Research*, 30(11), 2478-2483, (2002)

Gurevich A., Saveliev V., Vyahhi N., and Tesler G., QUAST: quality assessment tool for genome assemblies, Bioinformatics, 29(8): 1072-5, (2013)