Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (B)

Quality Control for RNA-Seq (QuaCRS): An Integrated Quality Control Pipeline

Karl W. Kroll[1,*], Nima E. Mokaram[2,*], Alexander R. Pelletier[1,*], David E. Frankhouser[2], Maximillian S. Westphal[2], Paige A. Stump[2], Cameron L. Stump[2], Ralf Bundschuh[1,3–5], James S. Blachly[1,§], and Pearlly Yan[1,2,§]

[1]Department of Internal Medicine, Division of Hematology, Ohio State University Comprehensive Cancer Center, Columbus, OH, USA. [2]Shared Genomics Resource, Ohio State University Comprehensive Cancer Center, Columbus, OH, USA. [3]Department of Physics, The Ohio State University, Columbus, OH, USA. [4]Department of Chemistry and Biochemistry, The Ohio State University, Columbus, OH, USA. [5]Center for RNA Biology, The Ohio State University, Columbus, OH, USA. *These three first authors contributed equally to this work. §These two corresponding authors contributed equally to this work.

ABSTRACT: QuaCRS (*Qua*lity *C*ontrol for *R*NA-*S*eq) is an integrated, simplified quality control (QC) system for RNA-seq data that allows easy execution of several open-source QC tools, aggregation of their output, and the ability to quickly identify quality issues by performing meta-analyses on QC metrics across large numbers of samples in different studies. It comprises two main sections. First is the QC Pack wrapper, which executes three QC tools: FastQC, RNA-SeQC, and selected functions from RSeQC. Combining these three tools into one wrapper provides increased ease of use and provides a much more complete view of sample data quality than any individual tool. Second is the QC database, which displays the resulting metrics in a user-friendly web interface. It was designed to allow users with less computational experience to easily generate and view QC information for their data, to investigate individual samples and aggregate reports of sample groups, and to sort and search samples based on quality. The structure of the QuaCRS database is designed to enable expansion with additional tools and metrics in the future. The source code for not-for-profit use and a fully functional sample user interface with mock data are available at http://bioserv.mps.ohio-state.edu/QuaCRS/.

KEYWORDS: RNA-seq, quality control, database, FastQC, RNA-SeQC, RSeQC

SUPPLEMENT: Computational Advances in Cancer Informatics (B)

CITATION: Kroll et al. Quality Control for RNA-Seq (QuaCRS): An Integrated Quality Control Pipeline. *Cancer Informatics* 2014:13(S3) 7–14 doi: 10.4137/CIN.S14022.

RECEIVED: June 2, 2014. **RESUBMITTED:** July 31, 2014. **ACCEPTED FOR PUBLICATION:** August 1, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: This work was supported by the National Institutes of Health [CA016058, CA140158, CA101140 and CA102031] and by an allocation of computing time from the Ohio Supercomputer Center (www.osc.edu). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: james.blachly@osumc.edu; pearlly.yan@osumc.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties.

Introduction

Advances in next-generation sequencing (NGS) technology are driving the numeric scales used in this sector of research in polar opposites. The data output (in units of hundreds of gigabytes per sequencing run) and the size of the computation server (in units of terabytes) for analyses are ever increasing, while the amount of input material used for sequencing library generation (in units of picogram or less) continue to decrease. For example, the SMART-seq2 technology described by Picelli et al.[1] enables sequencing even from a single cell. The speed of these changes is so rapid that monitoring their impact on the resultant data quality becomes paramount. Herein, we describe the development and application of an RNA-seq data quality control (QC) workflow called QuaCRS (*Qua*lity *C*ontrol for *R*NA-*S*eq) that leverages open-source RNA-seq QC tools as infrastructure and a MySQL database for data querying and for report generation via a graphical interface. QuaCRS is well

suited for NGS cores and research groups wanting to track RNA-seq data quality across time, personnel, studies, and library generation protocols/kits. Multiple user accounts can be created with defined access rights. Once established, the graphical user interface (GUI) will allow biologists and core personnel to search, filter, and output selected information to generate custom QC reports without the assistance of computation team members.

Motivations for QuaCRS: Instances of Errors, Biases, and Metrics for their Identification

RNA-seq (transcript profiling by high-throughput sequencing from stranded or non-stranded polyA+ RNAs or from rRNA-subtracted coding and non-coding RNAs) is a rapidly evolving technology with many dependencies, and therefore potential error sources.[2–5] Although the cost of data generation is not trivial, analysis costs (personnel as well as computation time) are far more extensive, and increasing efficiencies here would be of enormous benefit. Furthermore, mixing unreliable data into good quality data introduces variability that can mask the underlying biological effects being investigated. Even if the poor quality data are subsequently identified, reanalysis delays progression of projects and publication of data. This highlights the need to perform robust QC analysis of newly generated data as quickly as possible, using tools with metrics that reveal diverse systematic or sample-associated errors. Timely discovery of problematic data also allows core facility personnel and researchers to troubleshoot and optimize/modify laboratory protocols, resulting in earlier improvement of subsequent experiments and significant savings in time, money, and samples.

Sequencing instruments have on-board real-time data quality displays. For Illumina sequencers, information including the total number of raw reads, passed filter reads, Phred quality score, and estimation of reads associated with each sample barcode are provided. Additional quality parameters are available after signal processing using the Illumina CASAVA pipeline. This includes actual reads associated with sample barcodes as well as undemultiplexed reads and their associated barcodes.

The next level of RNA-seq data QC analysis involves expanded general as well as specific characteristics of sequencing reads, both pre- and post-alignment to the transcriptome. A survey of open-access information (NGS facility websites, publications, and meeting presentation slides) reveals that not all NGS facilities provide specific information about whether they perform QC analyses on RNA-seq data or what QC packages are used if they do so. For example, the Oxford Genomics Centre[6] indicates that their users can expect a primary QC report containing general QC metrics for each sample sequenced, and a secondary QC report containing QC metrics to assess the quality and specificity of the reads for transcriptome profiling and discovery, but whether they use open-source, proprietary, or custom software to generate these primary and the secondary QC reports is not mentioned. Where such information is provided, FastQC,[7] one of the pre-alignment QC packages, is by far the most often used software for QC analysis. FastQC, as the name implies, requires little time. This simple workflow provides informative metrics such as sequence duplication levels, overrepresented sequences, and *k*-mer content quickly, as it only requires pre-alignment data. As per-run sequencing data output increases, the computation time needed for aligning RNA-seq data increases dramatically because of the complexity of aligning reads across splice junctions. FastQC permits users to check data quality shortly after a run ends for metrics that shed light on RNA-seq data quality.

As mentioned, the diversity of RNA-seq protocols with varying systematic errors and potential batch effects necessitates QC evaluations of the more informative post-aligned data using software such as RNA-SeQC[8] and RSeQC.[9] RSeQC is often used for RNA-seq analysis, as evidenced by over 5000 downloads since its release in 2012.[10,11] Along with FastQC, it is wrapped into the Duke biopipeline[12] for RNA-seq preprocessing, alignment, and QC. The Ludwig Institute for Cancer Research in Stockholm and the University of New South Wales in Sydney also value RSeQC as a QC tool.[13,14] It is less clear how often RNA-SeQC is being used by the NGS community as it is less often referred to directly. However, as RNA-SeQC is a module in the GenePattern analysis platform,[15] groups that use GenePattern in their analysis workflow will likely be using this tool for QC analysis of their RNA-seq. The Mayo Clinic utilizes all three of these QC tools, stressing the importance of QC before pursuing downstream analysis.[16]

The many differences between current sample and library preparation methods imply that absolute cutoffs on QC parameters are meaningless and that aberrations must be detected in the context of input material and library preparations of similar types. In addition, accumulating QC data allows NGS core facilities to track data performance and explore unusual patterns that arise in a dataset through comparisons to valuable, historical data and cross-study meta-analyses. Anomalies may be exclusive to a sequencing run, a particular batch of samples, a tissue or data type, or a function of other QC metrics. A database is an ideal tool for probing data to reveal such factors. The questions at hand can easily be answered by condensing QC data with a few queries and filters. Currently, there are no documented tools that archive QC data for RNA-Seq in this manner. While we designed QuaCRS for our own use and find it to be enormously helpful, publication of this manuscript will allow us to share this and future versions of QuaCRS with the NGS community.

Description of the QuaCRS Pipeline

RNA-seq data preprocessing. QuaCRS is based on a workflow currently in place at The Ohio State University

Comprehensive Cancer Center Genomics Shared Resource NGS Core. (This portion of the description is not part of the QuaCRS pipeline, but is provided to illustrate steps upstream of the database for the purpose of continuity.) Our RNA-seq libraries are sequenced on an Illumina HiSeq 2500 sequencer. Signal intensities derived from the sequencer are converted into fastq files and demultiplexed into reads per sample using the Illumina CASAVA pipeline. These reads are trimmed for adapters, then aligned to known mitochondrial and ribosomal sequences (often present at varying amounts in total RNA-seq protocols, such as the Illumina Ribo-Zero rRNA reduction chemistry) to remove high abundance species that could bias gene expression analysis. The resulting reads are then mapped to the genome of interest using the STAR aligner.[17] Data then enter the QuaCRS pipeline described in the following sections, which include configuring input data, implementing QC tools, constructing the MySQL database, and generating composite QC reports from QuaCRS.

System requirements. QuaCRS can be run on a variety of systems with a basic requirement of a dual-core, 2 GHz processor and 8 GB of RAM. The QuaCRS database requires ~2 MB per sample file. Storage space and runtime vary greatly depending on the size and the number of data files included in each run. A sample file with 20 million 50 bp single-end reads takes ~90 minutes to complete on a quad-core workstation with 8 GB of RAM. A sample file with 70 million 100 bp paired-end reads will take eight hours to complete on a high performance computing cluster using 7 cores and 30 GB of RAM. Before launching the QuaCRS pipeline, the three required QC tools must be installed, complete with all relevant dependencies. Additional tools needed for QuaCRS include Picard tools,[18] SAMtools,[19] MySQL,[20] PHP,[21] ImageMagic,[22] and Numpy.[23] Users will also need a reference annotation in FASTA, GTF, and BED formats for gene body calculations in both post-alignment tools. Links to download these tools can be found in the installation readme on the QuaCRS website.

Input. The entry point to QuaCRS is a wrapper called QC Pack (see Fig. 1). It runs the three selected RNA-Seq QC tools currently used in our core: FastQC, RNA-SeQC, and RSeQC. This is also the entry point for both the obligatory metadata and other optional metadata deemed necessary for downstream data analyses. To launch the QuaCRS pipeline, a sample configuration file needs to be populated with the obligatory metadata, namely information needed to generate a searchable unique identifier for each sample in the QuaCRS database. This identifier is composed of the sample name, the study name, the start date of the sequencing run, and the lane number in which the sample was sequenced. Although no two samples would have the same combination of these four parameters, often the same sample from the same library preparation will be sequenced more than once (ie, different dates and lanes) for the purpose of achieving a predetermined number of sequenced reads. QuaCRS handles this scenario through another required field labeled "Run Description." This field marks data entries as combined if they represent data compiled from two or more sequencing runs that constitute the final data input for a sample, prior to downstream analyses. As such, these "combined" samples are not associated with date and lane information. The actual sequence read files for these "combined" samples are generated prior to their entry to the QC Pack, as QuaCRS does not perform the function of merging multiple sample data files. In addition to the obligatory metadata described above, other information important to downstream analyses but not to sample identification can also be input into QuaCRS. Examples of these optional metadata include total RNA extraction protocol, RNA-seq library generation protocol, quality metric of input total RNA for library generation (*R*NA *I*ntegrity *N*umber or RIN),[24] and sample barcode information. Once the QC Pack configuration file is populated, the QC Pack script is ready for the next step.

Quality control tools. Upon running the QC Pack script, FastQC is executed first as it is the least computationally demanding of the three QC tools. It accepts raw fastq files as input and generates sequencing QC metrics. Example metrics are as follows: per-base read quality, GC content, and per-base nucleotide bias. Unique to this QC tool, FastQC also identifies *k*-mers and adapter sequences in raw reads. Although evaluation of sequence data using FastQC will identify anomalies in the passed filter reads, it cannot reveal batch and systematic errors that aligned data can. RNA-SeQC and RSeQC accept aligned BAM files as input and are designed to assess mapping performance to give greater insight into data quality. For example, classifying aligned reads will provide exonic and intronic rates, gene body distribution, and strand specificity. RNA-SeQC and RSeQC share a number of similar QC functional checks; however, RSeQC has useful saturation functions such as junction saturation and RPKM saturation that RNA-SeQC is lacking. In addition, the two tools output their findings differently. RNA-SeQC summarizes the QC metrics in a table format, whereas RSeQC primarily generates graphical outputs. This makes the latter more suited for visual inspections while the former is more suited to querying and aggregating information across multiple samples. As we bundle QC parameters derived from multiple QC tools in QuaCRS, we allow end users to view QC metrics in more than one perspective. These parameters are complementary, and together can provide a much more useful overall picture than either alone. The presentation of "duplication rate" by RNA-SeQC and RSeQC illustrates this point well. RNA-SeQC reports read duplication rate as a single number for each sample. RSeQC outputs a plot that depicts the number of duplicated reads as a function of frequencies for both sequence duplicates and mapping duplicates. While the graphical output of RSeQC is more informative, it is also more difficult to summarize as a single metric.

There are other important computational differences between the two tools. RNA-SeQC has predetermined

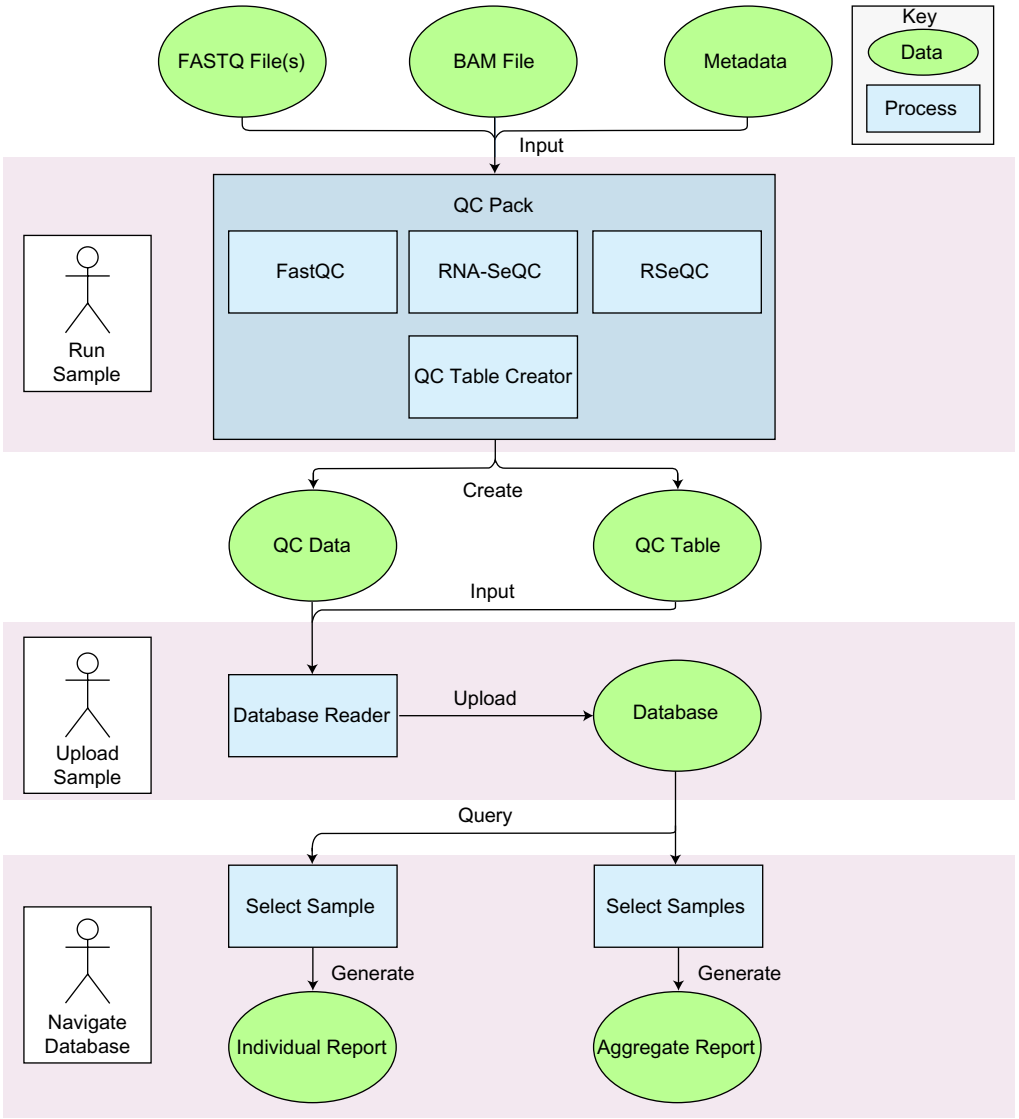

Figure 1. Workflow executed by QuaCRS. Raw data, aligned data, and additional metadata are provided as input to the QC Pack program. Executing QC Pack will run FastQC, RNA-SeQC, and RSeQC, and create a composite table containing the resulting metrics and image file paths. This table is then uploaded to the database using the database reader, after which it can be viewed with the web interface and for report generations based on individual samples or aggregates.

requirements for data input files. The QC Pack uses Picard tools and SAMtools to add read groups, sort the alignment file according to alignment locations by chromosome, mark duplicate reads, and index the BAM file before sending the alignment file to RNA-SeQC workflow. In contrast, RSeQC will accept alignment output files from an RNA-seq aligner without further modifications. Another key difference pertains to the parallelized processing nature of RSeQC. These two features render RSeQC a much faster tool to execute. The current version of the QuaCRS database is designed to provide a rapid survey of RNA-seq QC metrics, and to be easy to build and deploy as a MySQL database. Given these intents, we have only included a subset of RSeQC functions to be run by the QC Pack. However, the table structure of the QuaCRS database is readily amenable to including the remaining functions in future updates.

QuaCRS will process RNA-seq data sequentially through the three QC tools: FastQC, RNA-SeQC, and RSeQC. The output files are then ready for query and report generation as shown in Figure 2. The output file from one tool will trigger the launch of the subsequent tool until all steps are completed in the workflow. At each step, it will check to see whether the output from that tool already exists, and will move on to the next step if an output file is found. This design element allows efficient updating and management of ongoing studies present in QuaCRS. It eliminates the likelihood of rerunning any QC tools unnecessarily when new samples are uploaded to an existing study. There exist scenarios whereby one might need to bypass this functionality, eg, when replacing improperly generated QC data from a previous run. In this case, an additional argument can be passed to force the program to rerun all QC tools, regardless of whether QC output files for that

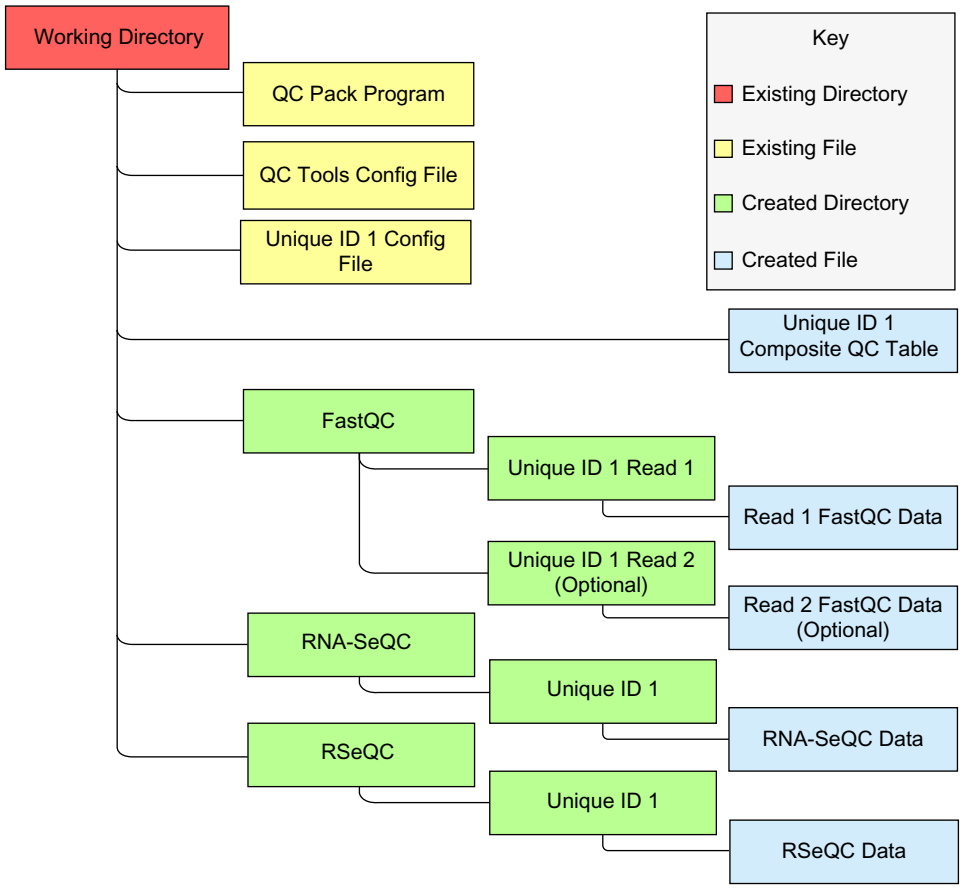

Figure 2. Directory structure produced and used by the QC Pack program. QC Pack runs on one sample at a time with a configuration text file passed as an argument. Before execution, this configuration file must be populated with the required metadata for the sample. In addition, a separate configuration file must be populated with the locations of the necessary tools for QuaCRS to run. These two configuration files, and the QC Pack program itself, are shown in yellow to indicate that they must exist before QC Pack is executed. Upon execution, the sample is given a unique identifier (ID) composed of several pieces of metadata. The workflow then creates the directories, shown in green, and populates them with data, shown in blue. It also produces a composite table of all QC metrics and images for the sample to be parsed into the database.

sample exists. At the completion of a run, a directory for each QC tool will appear. These directories will contain a subdirectory for each of the processed samples. Each sample subdirectory will contain a text file that will be parsed and copied into the database. QuaCRS provides the flexibility to upload samples to the database one at a time as their QC analyses finish, or upload the whole directory when all samples are processed.

Database structure. QuaCRS builds upon a MySQL database with a single, large table to organize the QC metrics and location identifiers of the QC image files. The database is populated by the launch of a Python program that parses the QC outputs and selects the pre-defined QC metrics and images as inputs to the QuaCRS data table. Many of the database fields may be left blank; however, any unexpected fields in the QC outputs will trigger an error and the sample will not be entered into the database. If all fields match to the predetermined metrics and image identifiers, the parsing program then executes the "create" or "update" command. These are distinct modes of operation. The "update" function is for samples that already have existing QC data in the database, eg, subsequent re-sequencing of the same sample library to increase read depth or addition of new metadata or additional QC parameters to already existing sample entries. The "create" function is for a new entry into the database. Providing these two distinct operations prevents the association of the QC information of a new sample with an existing sample in the case of a typographical error that causes a new sample name to be the same as an existing one.

Web interface and output. QuaCRS is designed with an interactive GUI as the front end. This is a user-friendly feature important to biologists and core personnel with limited computation experience. QuaCRS is also a time-saving tool for computation staff members, as it conveniently packages all the needed tools in one place, thereby simplifying the execution of a routine data analysis step. A fully functional version of the QuaCRS database populated with mock RNA-seq QC data (five samples, two of which are "combined"), its user interface, and a link for downloading the QuaCRS source code is available at http://bioserv.mps.ohio-state.edu/QuaCRS/. Information about how to log in, navigate, and query QC data using the web interface is provided in the "Readme" page on the QuaCRS website. QuaCRS users need prior administrative

approval before they can access their intended QC data. This security feature allows a single instance of QuaCRS to host projects owned by unrelated groups while providing controlled access to individuals designated by the data owners. As depicted in Figure 3, the main webpage displays a table view of all samples present in the database and accessible to the current user. This view facilitates examination of a single sample or multiple samples at once and for all or a few QC metrics. Database filtering can be applied via the "Columns" drop-down menu where QC check boxes are grouped into logical categories based on metric types. Checking and unchecking labels will show or hide the columns in the table view. The table can be sorted by clicking on the labels, and columns can be rearranged via drag and drop. This table may also be filtered by keywords using the search box. The selected keyword(s) can be searched across all columns, or in a specific column using a "column:value" format. The current search function works well to locate non-numeric values such as sample or study identifiers. In future versions, users will be able to filter numerical data by providing minimum or maximum values.

Users may investigate an individual sample in detail by clicking on the sample name in the table view. This view will navigate to a new page with all the QC output information associated with the selected sample, as shown in Figure 4. Data are grouped into blocks of similar metrics that can be displayed together. There is also a block for all the plots produced by FastQC and RSeQC, and each one can be individually enlarged. Users can download the entire QC information for a single sample with the "Download Report" button on this page. The download file will contain a two-row table featuring all the columns in the database.

It is possible to view and compare multiple samples at once by generating an aggregate report. This can be achieved from the table view by selecting the desired samples and clicking the "Aggregate" button. This will generate a detailed view similar to the individual sample view shown in Figure 4. An important difference is that for every quantitative metric, the aggregate report will show the minimum, maximum, and average values of the selected sample set instead of a single value. The qualitative FastQC metrics are summarized as the counts of selected samples that "passed", "failed," or generated a "warning". The "plots" menu is not available in this view, but the distribution of each quantitative metric is provided graphically as a box plot in an adjacent column. As in the single view option, aggregate results can be downloaded as a report.

Summary

Here, we describe and provide a new database approach for RNA-seq QC analyses and report generation. Our tool, QuaCRS, allows a core facility or a large laboratory to store and query QC metrics and sample-related metadata across multiple samples and studies to track library generation processes, robustness of different RNA-seq library kits, and sequencing run qualities. This allows core and laboratory personnel to identify problematic samples, extraction procedures, and sequencing runs for process optimization. QuaCRS is user-friendly in that it wraps three open-source QC tools, stores their results in a MySQL database, and allows the results to be queried through a web-based front end. Once established, QuaCRS can be executed by laboratory personnel after brief command line level training. Ultimately, the advantage of implementing this workflow over running the three QC tools separately is the query-able nature of the database structure. This offers users the ability to track quality and performance across large numbers of samples, which would be far more difficult without the flexibility of the database approach.

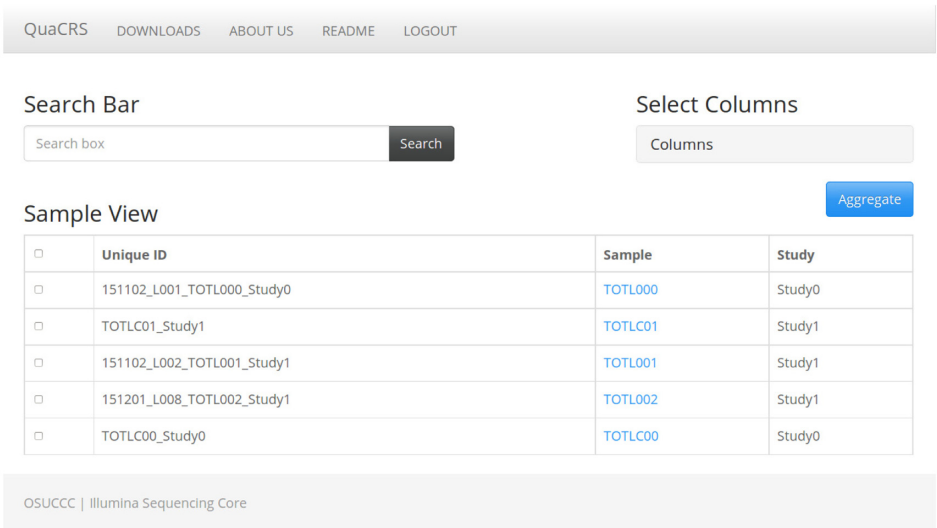

Figure 3. Sample View. Columns can be shown or hidden through the "Select Columns" menu. Visible columns can be sorted by clicking the header and reordered with drag and drop. Clicking on a blue sample name will show a sample report for that sample. Clicking the "Aggregate" button will generate an aggregate report for the samples currently selected using the check boxes.

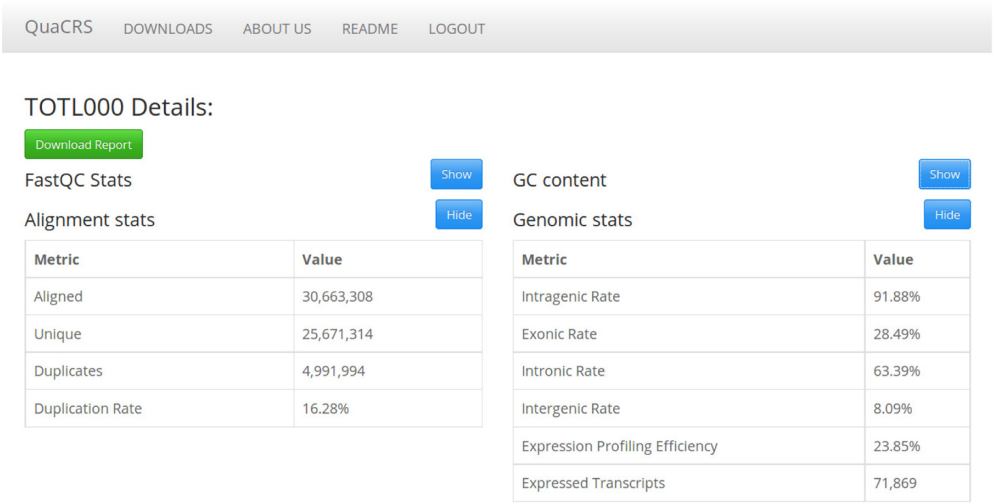

Figure 4. Sample report. This is a detailed view for a specific sample in the database. QC metrics are grouped into logical segments that can be individually shown or hidden by clicking the "Show/Hide" buttons. Qualitative FastQC metrics are color-coded according to whether they pass, fail, or receive a warning. Quantitative metrics are reported numerically or as percentages, depending on the type of metric. This view also has a drop-down menu to show or hide all plots generated by FastQC and RSeQC (not shown). Aggregate reports supply similar information, but for multiple samples. In an aggregate report, qualitative metrics from FastQC are represented as the number of selected samples that passed, failed, or generated warnings. Quantitative metrics are represented using minimum, maximum, and average values for the selected samples, as well as box plots summarizing the data across the selected samples.

Already, QuaCRS has streamlined the QC analyses and automated custom report generation in our core facility. By querying QC metrics across studies and samples, we noted that "expression profiling efficiency" works in concert with "estimated library size" and "duplication rate" (metrics from RNA-SeQC) to provide meaningful information on sample quality, such as the effects of incomplete rRNA removal in transcriptome profiling. Using QuaCRS, we also observed that in RNA-seq data derived from low-quality samples, exonic rates are greatly altered, while intronic and intergenic rates are much less perturbed. The greatly enhanced functionality and ease of use provided by QuaCRS versus existing tools for QC metrics is increasingly important as we accelerate our sequencing rate and sample diversity, and also as we adopt multiple new library generation approaches for RNA-seq analyses. It is our intent to further expand the search functionalities of QuaCRS and to include more QC metrics from RSeQC outputs. These features will be released in stages in future revisions of QuaCRS. With the continuous advancement of global transcriptomic analyses, the development and refinement of database approaches such as QuaCRS is critical in ascertaining that only high-quality data are used for downstream analyses.

Acknowledgements

The authors would like to acknowledge the help in library generation by John Curfman as well as helpful discussions regarding the development of the QuaCRS workflow by Javkhlan-Ochir Ganbat and Michael W. Zoller. The authors would also like to acknowledge Dr. David Lucas, Department of Internal Medicine, for his assistance in this manuscript.

Author Contributions

Conception of idea: KWK, PY, ARP, NEM, RB, JSB, DEF. Preparation of manuscript: KWK, PY, ARP, NEM, RB, JSB, DEF. Creation of QC Pack: KWK, ARP. Construction of MySQL database: NEM, ARP, KWK. Preparation of readme document: KWK, NEM, ARP. Creation of workflow: KWK, PY, DEF, MSW, ARP, CLS, PAS. RNA-seq library preparation and sequencing: PAS, CLS, PY. Metadata information: PAS, CLS, PY. Overall coordination: RB, JSB, PY. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Picelli S, Faridani OR, Bjorklund A, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*. 2014;9:171–81.
2. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11:733–9.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5:621–8.
4. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. *Genome Res*. 2011;21:2213–23.
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;19:57–63.
6. RNA-Seq Analysis. Oxford Genomics Center Web site. http://www.well.ox.ac.uk/ogc/rna-seq/.
7. FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics Web site. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
8. DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*. 2012;28:1530–2.
9. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*. 2012;28:2184–5.
10. RSeQC RNA-seq data QC. Sourceforge Web site. http://sourceforge.net/projects/rseqc/files/.
11. Rseqc RNA-seq quality control package. Google Code Web site. https://code.google.com/p/rseqc/downloads/list.
12. Bioinformatics Analysis Pipeline. Duke Center for Human Genome Variation Web site. http://redmine2.chgv.lsrc.duke.edu/projects/biopipeline/wiki/RNA-seq.

13. Hupe M, Li MX, Gillner KG, Adams RH, Stenman JM. Evaluation of TRAP-sequencing technology with a versatile conditional mouse model. *Nucleic Acids Res*. 2014;42(2):e14.
14. Mills JD, Kavanagh T, Kim WS, et al. Unique transcriptome patterns of the white and grey matter corroborate structural and functional heterogeneity in the human frontal lobe. *PLoS One*. 2013;8(10):e78480.
15. RNA-seq QC in GenePattern. GenePattern Web site. http://www.broadinstitute.org/cancer/software/genepattern/gp_guides/indeptharticles/sections/RNAseqQCinGP.
16. Nair A, Hart S, Sicotte H, et al. Optimizing sequence yield & interpreting data quality for RNA-Seq. Poster session presented at: International Society for Computational Biology. 20th Annual International Conference on Intelligent Systems for Molecular Biology; July 15–17, 2012; Long Beach, CA.
17. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
18. Picard. Sourceforge Web site. http://picard.sourceforge.net/.
19. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinforma Oxf Engl*. 2009;25:2078–9.
20. MySQL. MySQL Web site. http://dev.mysql.com/.
21. PHP. PHP Web site. http://us2.php.net/.
22. ImageMagick Convert Command-line Tool. ImageMagick Web site. http://www.imagemagick.org/script/convert.php.
23. NumPy Reference. SciPy Web site. http://docs.scipy.org/doc/numpy/reference/.
24. Schroeder A, Mueller O, Stocker S, et al. The RIN: an RNA integrity number for assessing integrity values to RNA measurements. *BMC Mol Biol*. 2006;7:3.