HTSeqQC: **A Flexible and One-Step Quality Control Software for High-throughput Sequence Data Analysis**

Renesh Bedre[1], Carlos Avila[2], Kranthi Mandadi[1,3*]

[1]Texas A&M AgriLife Research and Extension Center, Texas A&M University, Weslaco, TX, USA
[2]Department of Horticultural Science, Texas A&M University, College Station, TX, USA
[3]Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, USA

*corresponding author: kkmandadi@tamu.edu

Abstract

Motivation: Use of high-throughput sequencing (HTS) has become indispensable in life science research. Raw HTS data contains several sequencing artifacts, and as a first step it is imperative to remove the artifacts for reliable downstream bioinformatics analysis. Although there are multiple stand-alone tools available that can perform the various quality control steps separately, availability of an integrated tool that can allow one-step, automated quality control analysis of HTS datasets will significantly enhance handling large number of samples parallelly.

Results: Here, we developed HTSeqQC, a stand-alone, flexible, and easy-to-use software for one-step quality control analysis of raw HTS data. HTSeqQC can evaluate HTS data quality and perform filtering and trimming analysis in a single run. We evaluated the performance of HTSeqQC for conducting batch analysis of HTS datasets with 322 sample datasets with an average ~ 1M (paired end) sequence reads per sample. HTSeqQC accomplished the QC analysis in ~3 hours in distributed mode and ~31 hours in shared mode, thus underscoring its utility and robust performance.

Availability and implementation: HTSeqQC software, Docker image and Nextflow template are available for download at https://github.com/reneshbedre/HTSeqQC and graphical user interface (GUI) is available at CyVerse Discovery Environment (DE) (https://cyverse.org/). Documentation available at https://reneshbedre.github.io/blog/htseqqc.html and https://cyverse-htseqqc-cyverse-tutorial.readthedocs-hosted.com/en/latest/ (for CyVerse).

Contact: Kranthi Mandadi (kkmandadi@tamu.edu)

Supplementary information: Supplementary information provided in Supplementary File 1.

Keywords: High-throughput sequencing (HTS), Quality control, Parallel computing, CyVerse, Nextflow

Introduction

Advancements in high throughput sequencing (HTS) technologies transformed biological research. HTS largely replaced conventional low-throughput Sanger-based sequencing technologies for genome-scale studies. Multiple genome sequencing approaches (DNA-seq, RAD-seq, GBS, AgSeq) are being used to study genetic variations, discovery of novel genes, high-throughput genotyping, biomarker discovery, and precision medicine (Awika, et al., 2019; Bolger, et al., 2014; Edwards and Batley, 2010; Kandel, et al., 2019; Suwinski, et al., 2019). Similarly, transcriptome-level sequencing approaches (RNA-seq) allows determining the steady-state expression of genes, identification of novel transcripts and isoforms, alternatively splicing patterns, polymorphisms, gene co-expression networks, allele-specific expressions, and long non-coding RNAs (lincRNA) (Bedre, et al., 2019; Bedre, et al., 2019; Xu, et al., 2018; Zhou, et al., 2018). Illumina's benchtop and production-scale sequencers, which are by far the most widely used HTS platforms, can generate up to 1 and 20 billion sequence reads per run, respectively. This sequencing output is expected to rapidly increase due to further advancements

in the Illumina technologies. As such advanced bioinformatics tools are necessary to accelerate the analysis and efficiently manage the large volumes of data generated by the Illumina sequencing platforms.

In all HTS experiments, a critical first step is raw data quality assessment, filtering and trimming. HTS raw data often contains sequences of poor quality along with adapter or primer contaminations, and uncalled bases (N), which if not removed can significantly hamper the downstream bioinformatics analysis leading erroneous conclusions (Conesa, et al., 2016; Trivedi, et al., 2014; Zhou, et al., 2013). Several standalone software tools are available for quality control analysis, trimming, and filtering of HTS data such as FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), NGS QC (Patel and Jain, 2012), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), NGS QCbox (Katta, et al., 2015), Trimmomatic (Bolger, et al., 2014), and QC-Chain (Zhou, et al., 2013). However, most of them have limitations. For instance, FastQC performs only quality check of the data, and does not filter or trim of raw sequences (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Conversely, FASTX-Toolkit although is equipped to perform quality filtering does not support parallel computing to handle large-scale batch analysis. Other software tools such as NGS QC, QC-Chain, Trimmomatic, and NGS QCbox have limited features for quality filtering, handles few samples at a time, dependent on other open-source software tools, and have a need to be run separately for different quality control features. With HTS becoming more accessible and affordable, it is crucial to develop a flexible and integrative tool that can not only perform thorough quality control analysis, but can handle several hundred samples parallelly, with the least number of data handling steps.

Here, we present HTSeqQC, which is an open-source and easy-to-use quality control analysis software tool for cleaning raw HTS datasets generated from Illumina sequencing platforms. HTSeqQC is a flexible, one-step quality control software tool and can handle large number of samples. HTSeqQC integrates filtering and trimming modules for single and paired end HTS data and supports parallel computing for batch quality control analysis. In addition to the quality filtering and trimming analysis, HTSeqQC generates statistical summaries and visualization to assess the quality of the HTS datasets. HTSeqQC can be used as command-line interface (CLI) as well as GUI. The GUI is available through CyVerse DE (https://cyverse.org/).

Implementation

HTSeqQC is a standalone open-source command-line software tool developed using Python 3 for quality control analysis of HTS data generated from Illumina sequencing platforms. At its core, HTSeqQC consists of two main modules that are intended to filter and trim single and paired-end sequence datasets. Both modules were implemented using parallel computation to increase the performance of quality control analysis by allocating the input workload to the multiple CPUs. By default, there are only two CPUs, however this number can be changed as per user preferences. To some extent, HTSeqQC works similarly to MapReduce where it splits the large sequence file into smaller chunks, distributes the input data to multiple CPUs, and combines the input from each process to produce a final output.

HTSeqQC checks quality issues in the raw HTS datasets and performs the quality filtering and/or trimming in a single run for removing low-quality bases, adapter contamination, and uncalled bases (N) as per the settings of the user. The entire quality control can be completed in a single input command. By default, HTSeqQC filters out the sequence reads with Phred quality score ≤ 20. We did not add the feature of duplicate reads removal in HTSeqQC as this feature is directly related to the gene quantification and estimation of expression (Zhou, et al., 2018). A flowchart HTSeqQC analysis is shown in Figure 1.

HTSeqQC can handle batch analysis of multiple sequencing datasets at a time. It accepts FASTQ file format as input and produces FASTQ or FASTA file format as output. HTSeqQC also generates summary statistics and visualization outputs for the filtered cleaned HTS datasets. All the outputs by default are saved in the same directory containing the raw RNA-seq input datasets.

HTSeqQC was primarily designed to run on the Linux and Mac operating systems as command-line interface CLI (Figure 2), however, it can also run on the Windows operating systems using virtual machine. In addition, we made HTSeqQC publicly available as GUI (Figure 2) on CyVerse (Goff, et al., 2011; Merchant, et al., 2016), which would be useful for experimental biologists with minimal bioinformatics training. CyVerse is an open-access, scalable, comprehensive, data analysis and management infrastructure for large-scale data analysis developed for life science research. CyVerse can be accessed through world-wide-web, and users can register for free at https://user.cyverse.org/register. To use the HTSeqQC as GUI on CyVerse, it is necessary to have CyVerse account. Additionally, we have also provided the docker image and Nextflow template for running the HTSeqQC.

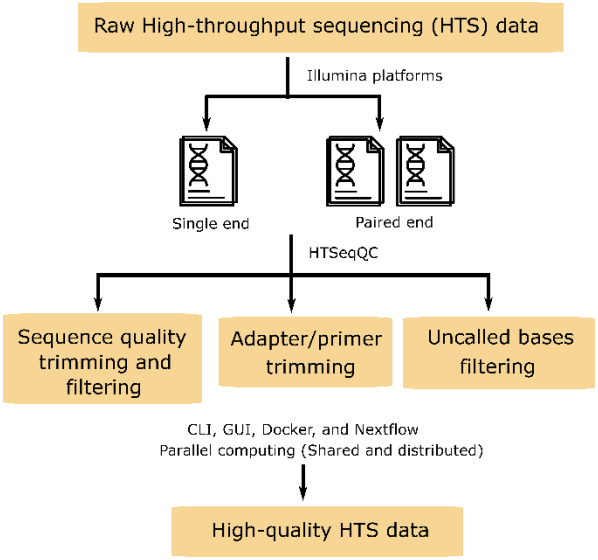

Figure 1: **Flowchart of HTSeqQC analysis.** HTSeqQC includes two main modules for quality control analysis of single and paired-end HTS datasets generated from Illumina sequencing platforms. HTSeqQC filters and trims the raw HTS datasets to remove low-quality bases, adapter or primer contamination, and uncalled bases to generate high-quality datasets for downstream bioinformatics analysis.

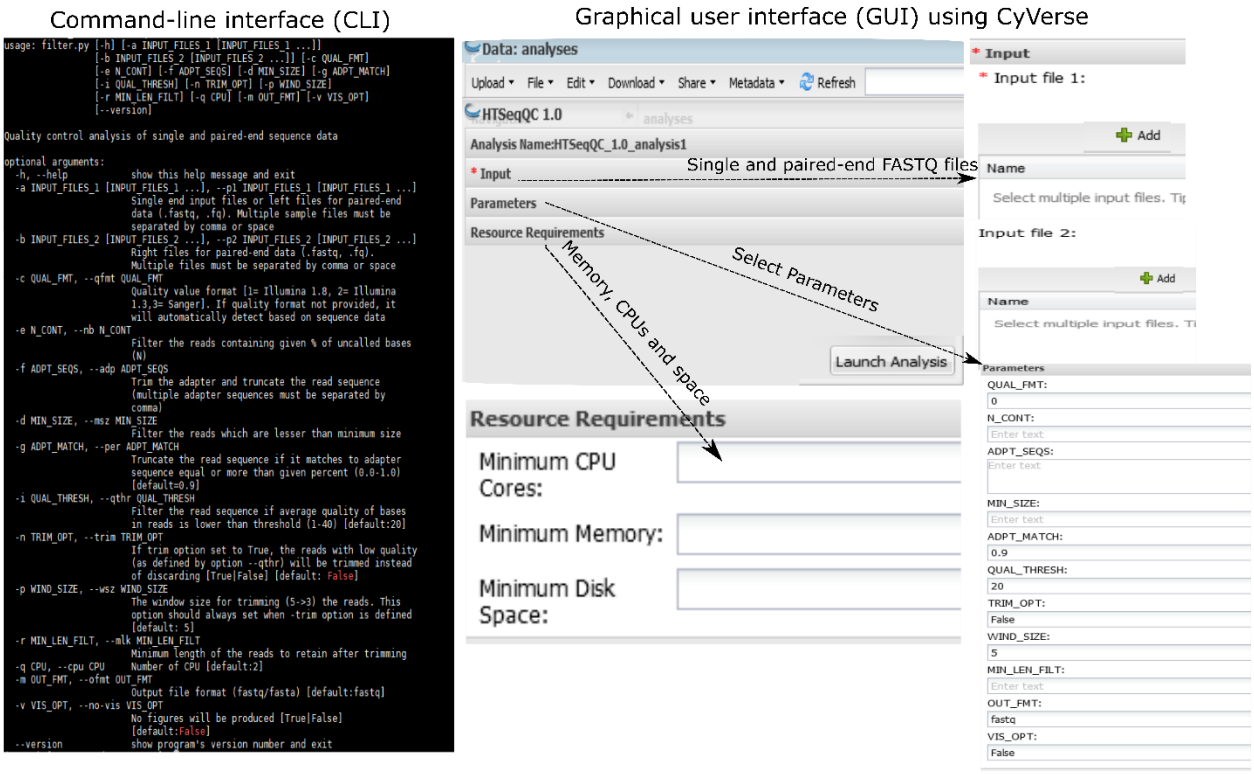

Figure 2: Command-line interface (CLI) and Graphic-user interface (GUI) of HTSeqQC. HTSeqQC can be launched as CLI or GUI modules. For CLI, the HTSeqQC need to installed on a local system, whereas GUI is preinstalled and ready to use online at CyVerse (https://cyverse.org/).

Case studies

To evaluate the performance of HTSeqQC for quality control analysis, we analyzed several datasets using a 64 GB RAM and 20 CPUs computing node. First, we analyzed a single-end raw RNA-seq dataset generated on Illumina HiSeq 2000 platform corresponding to cotton, a dicot plant (BioProject accession PRJNA275482 and SRA ID: SRR1805340, Table 1) (Bedre, 2016; Bedre, et al., 2015). HTSeqQC analysis was performed with the default parameters. HTSeqQC automatically detected the Illumina sequence quality variant and filtered out the reads with Phred quality score ≤ 20. In total, HTSeqQC filtered out ~5 M reads (Figure 3 and Supplementary File 1A).

Second, we evaluated multiple paired-end raw RNA-seq datasets generated on Illumina HiScanSQ platform corresponding to sugarcane, a monocot plant (BioProject accession PRJNA291816 and SRA IDs: SRR2165176, SRR2165177, SRR2165178) (Bedre, et al., 2019; Schaker, et al., 2016). Initially, we ran HTSeqQC on one paired-end dataset (SRR2165176) using default parameters (Table 1). HTSeqQC was able to analyze the Illumina sequence quality variants and filtered out the reads with Phred quality score ≤ 20 (Supplementary File 1B). For instance, the HTSeqQC filtered out the ~250 K sequence reads which were below the quality threshold (Supplementary File 1B). Next, we ran the HTSeqQC with customized parameters for filtering adapter sequences, quality thresholds, and uncalled bases. In total, HTSeqQC filtered

out ~451 K reads and trimmed ~20 K reads (Supplementary File 1C). All the HTSeqQC commands used for performing the above analyses are provided in the README file.

Table 1: Summary of the quality control analysis of single and paired-end datasets with single and multiple samples performed using HTSeqQC CLI.

SRA Accession or Samples	Read Type (# Samples)	# Sequence reads (File size in GB[a])	Read length (bp)	# CPUs (Parallel computing)	Run time (Min)
SRR2165176	Paired (1)	8,583,424×2 (5.6)	100	18 (Shared)	9
SRR2165176 SRR2165177 SRR2165178	Paired (3)	8,583,424×2 (5.6) 9,282,222×2 (6) 9,918,081×2 (6.4)	100 100 100	18 (Shared)	28
SRR1805340	Single (1)	82,059,811 (27)	100	18 (Shared)	40
Tomato GBS[b]	Paired (322)	~ 1,000,000×2 (206)	150	18 (Shared) 18 (Distributed)	1855 157

[a]Combined file size of two files for paired-end sequence reads
[b]322 tomato genotypes were sequenced using low-coverage whole genome sequencing by Illumina HiSeq 4000 (Kandel, et al., 2019). The number of sequences reads are average of all the datasets.

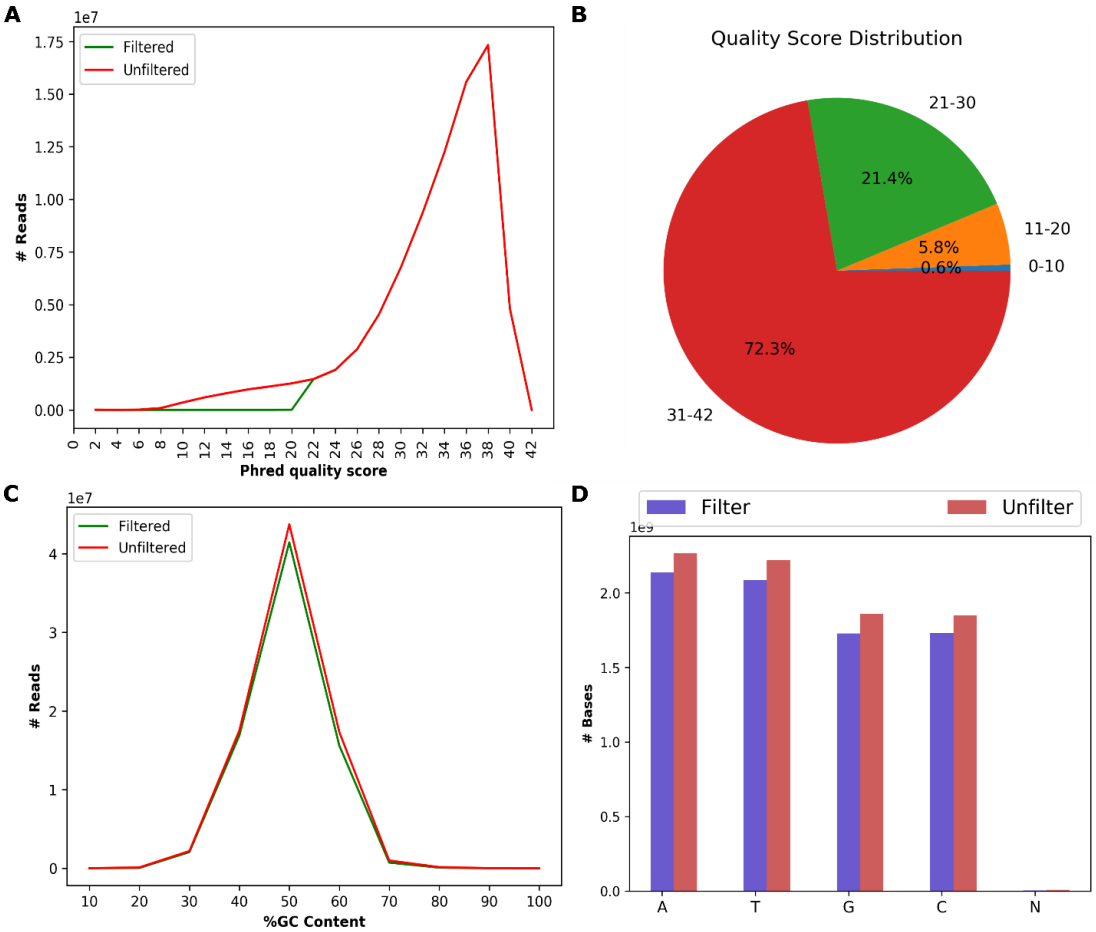

Figure 3: The sequence quality evalaution of HTS datatsets performed by HTSeqQC. A) The sequence quality distribution among the raw (unfiltered) and cleaned (filtered) sequence data, B) The distribution of percenatges of sequence reads with quality score, C) The percentage

GC content distribution among the raw and cleaned sequence data, D) The content of nucleotide bases in raw and cleaned sequence data.

In addition to quality control analysis, we also evaluated the processing time and batch handling of HTSeqQC using default parameters. HTSeqQC took ~9 min to perform the quality filtering analysis of ~9 M paired-end sequence reads, and ~40 min for ~82 M single-end sequence reads with 18 CPUs (Table 1).

Lastly, we analyzed 322 paired-end genotyping-by-sequencing (GBS) datasets corresponding to tomato (Kandel, et al., 2019). These datasets had ~1M (×2) sequence reads per sample. For this experiment, we also compared parallel computing in shared vs. distributed mode with 18 CPUs per computing node using Nextflow. We were able to analyze the 322 datasets in ~1855 min (~31 hours) and ~157 min (~3 hours) in the shared and distributed computing modes, respectively (Table 1).

Advantages of HTSeqQC over existing quality control analysis software

There are multiple quality control analysis tools such as the FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), NGS QC (Patel and Jain, 2012), QC Chain (Zhou, et al., 2013), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and NGS QCbox (Katta, et al., 2015) on various important quality control features (Table 2). However, they are not as integrated as HTSeqQC (Table 2). For instance, programs such as FastQC only provides the quality control statistics of the raw HTS data and cannot be used for quality filtering. The FASTX-Toolkit does quality filtering, but the various trimming, filtering options are not integrated, and each analysis must be performed separately. The NGS QCbox comes close, however only allows sequence read trimming based on a quality threshold for specific window size and it is highly dependent on other open-source software tools for quality control analysis (Katta, et al., 2015). . NGS QC allows filtering and trimming of low-quality reads similar to FASTX-Toolkit, however these software tools are independent and must be run separately.

Table 2: The key sequence quality features comparison of HTSeqQC with other leading and equivalent quality control software tools.

Features	**HTSeqQC**	**FastQC[a]**	**NGS QC**	**QC-Chain**	**FASTX-Toolkit**	**NGS QCbox**
Low quality filtering	Yes	Yes	Yes	Yes	Yes	No
Uncalled bases (N)	Yes	No	Yes*	No	No	No
Adapter or primer trimming	Yes	Yes	Yes	Yes	Yes	No
Multiple adapter support\	Yes	Yes	Yes	Yes	No	No
Minimum read length filtering	Yes	No	No	No	No	Yes
Low quality trimming	Yes	No	Yes*	Yes	No	Yes

Paired-end support	Yes	No	Yes	Yes	No	Yes
Paired-end sequence order	Yes	No	Yes	Yes	No	Yes
Visualization	Yes	Yes	Yes	Yes	Yes	Yes
Parallel computing	Yes	Yes	Yes	Yes	No	Yes
Automatic quality variant detection	Yes	No	Yes	No	No	No
FASTA Output	Yes	No	No	No	No	No
Programming	Python 3	Java	PERL	C++	C	C
GZIP FASTQ input	Yes	No	Yes	No	No	Yes
Multiple sample support (batch analysis)	Yes	Yes	Yes	No	No	Yes
GitHub	Yes	Yes	No	No	Yes	Yes

[a]FastQC only provides quality metric and does not filter or trim HTS data
*Separate tools for given features

The removal of sequence reads containing uncalled bases (N) is critical to improve the accuracy of the sequence reads (Minoche, et al., 2011). HTSeqQC has this feature of filtering the sequence reads based on the content of the uncalled bases (N). Although NGS QC allows this analysis, it is not integrated, and must run separately (Patel and Jain, 2012). In addition to overcoming several of these limitations, HTSeqQC supports common features with NGS QC and NGS QCbox such as allowing inputs in the compressed GZIP file format for seamless input of large datasets (Table 2). Furthermore, similar to NGS QC, HTSeqQC has an integrated function to automatically detect the FASTQ quality variants (Table 2). Lastly, only HTSeqQC provides a parameter to output the cleaned FASTQ data in FASTA format, avoiding using additional file format conversion tools. A limitation of HTSeqQC is that this version (v1.0) was developed specifically for Illumina generated FASTQ datasets, however, it could be utilized with other sequencing platforms given the input is FASTQ and supports same quality formats as Illumina. We focused on Illumina mainly because it is the most-widely employed platform. However, we will continue development and subsequent versions of HTSeqQC will be released to expand its use to other platforms such as PacBio and/or Nanopore Sequencing.

Conclusion

HTSeqQC is an open-source, integrated, and easy-to-use software designed for one-step quality control visualization and quality filtering of raw HTS data generated using the Illumina sequencing platforms. The flexiblility to detect and remove the low-quality sequences, adapter or primer contamination, uncalled bases, in a single run, greatly enhances the automation of HTS data analysis projects. Because HTSeqQC can be implemented by parallel computing, it enables batch handling of large number (>300) of datasets. In addition to the command line interface, HTSeqQC is available as graphic user interface that is accessible freely through CyVerse. The later should significantly facilitate its use among biologists without much prior bioinformatics or command line computing experience.

Author contributions

RB developed the software, platform and conducted the analysis. CA, and KM supervised the study, data analysis and interpretation. All authors have read, reviewed, and approved the manuscript.

Acknowledgements

We thank Upendra Devisetty, Reetu Tuteja, and Sarah Roberts for their assistance with installing HTSeqQC at CyVerse DE, which was made possible through CyVerse's External Collaborative Partnership program. We acknowledge support of Louisiana State University Agricultural Center and Louisiana State University High Performance Computing resources (http://www.hpc.lsu.edu/) for supporting early stages of the HTSeqQC development as part of Ph.D. research of RB. We also acknowledge support of Texas A&M High Performance Research Computing Center (http://hprc.tamu.edu/) resources and sequencing support of the Texas A&M AgriLife Genomics and Bioinformatics Service (https://txgen.tamu.edu/). This work was supported in part by funds from Texas A&M AgriLife Research Insect-vectored Disease Seed Grant, Foundation for Food and Agricultural Research New Innovator Award (2018-534299) and USDA-NIFA-CDRE (2018-70016-28198) to KM.

References

Awika, H.O., *et al.* Developing Growth-Associated Molecular Markers Via High-Throughput Phenotyping in Spinach. *Plant Genome-Us* 2019;12(3).
Bedre, R. Genome-wide Transcriptome Analysis of Cotton (Gossypium hirsutum L.) to Identify Genes in Response to Aspergillus flavus Infection, and Development of RNA-Seq Data Analysis Pipeline. 2016.
Bedre, R., *et al.* New Era in Plant Alternative Splicing Analysis Enabled by Advances in High-Throughput Sequencing (HTS) Technologies. *Front Plant Sci* 2019;10(740).
Bedre, R., *et al.* Genome-wide alternative splicing landscapes modulated by biotrophic sugarcane smut pathogen. *Sci Rep* 2019;9(1):8876.
Bedre, R., *et al.* Genome-wide transcriptome analysis of cotton (*Gossypium hirsutum* L.) identifies candidate gene signatures in response to aflatoxin producing fungus *Aspergillus flavus*. *PLoS One* 2015;10(9):e0138025.
Bolger, A.M., Lohse, M. and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114-2120.
Bolger, M.E., *et al.* Plant genome sequencing - applications for crop improvement. *Curr. Opin. Biotechnol.* 2014;26:31-37.
Conesa, A., *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17:13.
Edwards, D. and Batley, J. Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.* 2010;8(1):2-9.
Goff, S.A., *et al.* The iPlant Collaborative: Cyberinfrastructure for plant biology. *Front Plant Sci* 2011;2(34).

Kandel, D.R., *et al.* Genetic Diversity and Population Structure of Tomato (Solanum lycopersicum) Germplasm Developed by Texas A&M Breeding Programs. *American Journal of Plant Sciences* 2019;10(7):1154-1180.
Katta, M.A., *et al.* NGS-QCbox and Raspberry for Parallel, Automated and Rapid Quality Control Analysis of Large-Scale Next Generation Sequencing (Illumina) Data. *PLoS One* 2015;10(10):e0139868.
Merchant, N., *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* 2016;14(1):e1002342.
Minoche, A.E., Dohm, J.C. and Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* 2011;12(11):R112.
Patel, R.K. and Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012;7(2):e30619.
Schaker, P.D., *et al.* RNAseq transcriptional profiling following whip development in sugarcane smut disease. *PLoS One* 2016;11(9):e0162237.
Suwinski, P., *et al.* Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front Genet* 2019;10:49.
Trivedi, U.H., *et al.* Quality control of next-generation sequencing data without a reference. *Front Genet* 2014;5:111.
Xu, W., *et al.* Differential expression networks and inheritance patterns of long non-coding RNAs in castor bean seeds. *Plant J.* 2018;95(2):324-340.
Zhou, Q., *et al.* RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. *BMC Genomics* 2018;19(1):144.
Zhou, Q., *et al.* QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One* 2013;8(4):e60234.