QuaC: Implementing Quality Control Best Practices for Genome Sequencing and Exome Sequencing Data

Manavalan Gajapathy, Brandon M. Wilk, Donna M. Brown, Elizabeth A. Worthey Center for Computational Genomics and Data Science, University of Alabama at Birmingham, Birmingham, AL

Introduction

- Quality Control (QC) of genome sequencing (GS) and exome sequencing (ES) data is necessary to ensure that data are of sufficient quality for downstream analyses. Examples of QC failures that need to be identified include poor sequencing data quality, data not matching to sequenced individual's expected metadata (sex, ancestry, relatedness), sample contamination, batch effects, etc.
- While several QC tools are available to perform QC at various levels post sequencing, their output needs to be reviewed and interpreted in a very manual process. This is a major challenge in large projects where standardization and consistency are highly desired in terms of QC metrics used and the threshold values utilized for such metrics.
- Further, logging the results of QC review and disseminating them with those involved in downstream analysis in an understandable format can be challenging. This could result in downstream users not utilizing the prior QC review results or re-reviewing them on their own and thereby wasting time and effort.

Results

- We have developed a pipeline called QuaC to run QC tools at the various stages of secondary analysis, to summarize QC results, and to allow visualization and sharing of results in an easy-to-consume format. We further standardized the logging of QC review results.
- QuaC runs several QC tools and further consumes QC results produced by its upstream companion alignment and small variant caller pipeline (Table 1).
- It includes a tool called QuaC-Watch, which consumes results from all the QC tools, performs QC checkup, and then summarizes whether samples have passed user-defined thresholds for QC metrics of interest or not. QuaC-Watch comes with pre-configured thresholds for both GS and ES data, and they were curated based on literature, in-house analyses and prior experience (Fig. 1).
- QuaC aggregates results produced by all the QC tools as well as QuaC-Watch using MultiQC, both at the single-sample- and project-level, and generates a single stand-alone, easy-to-distribute HTML report file (Fig. 1).
- We further devised a "Sample QC database" where manual QC review results would be stored using controlled flags (Table 2).

QuaC makes QC easy for Genome and Exome Sequencing data



Pipeline runs several QC tools for BAM and VCF files and accepts QC results of FASTQs.



QuaC-Watch tool performs QC checkup based on the expected, pre-configured thresholds and summarizes the results for easy consumption.



Aggregates and visualizes QC results using MultiQC, both at the single sample and project levels.

Multi⊕C	QuaC-V	Vatch	า				
v1.9 (fdbbd79)	This section con	tains Qua(C-Watch results	. QuaC-Wa	tch summarizes if sample	s have passed the (QC thresh
General Stats	Overall Q	uaC-V	Vatch Su	mmar	y		
QuaC-Watch	Overall QuaC-Wa	atch summ	nary of results f	rom severa	I QC tools		
Overall QuaC-Watch Summary	🗳 Copy table	III Configu	re Columns	Plot Sh	owing $^{3}/_{102}$ rows and $^{9}/_{9}$ colu	mns.	
FastQC (trimmed)	Sample Name	fastqc	qualimap_o	verall qu	ualimap_chromosome_s	pecific picard	picard_c
Qualimap - Overall stats		fail	pass	F	bass	fail	pass
Qualimap - Chromosome stats	1.00	fail	fail	ſ	ail	fail	pass
Picard		fail	pass	ſ	ail	fail	pass
Picard-dups	E 100 //		D.				
Bcftools stats	FastQC (t	rimme	ed)				
Variant frequency per contig	Quick summary	of FastQC	(trimmed) resu	ts. See Fas	tQC section below for de	tailed results.	
VerifyBAMID	Sopy table	🗰 Configu	re Columns	Plot Sh	owing $^{14}/_{572}$ rows and $^{10}/_{10}$ co	olumns.	
FastQ Screen (trimmed)	Sample Name	per_ba	se_sequence_	quality p	per_tile_sequence_qualit	ty per_sequence	_quality_
QualiMap	-R	1 pass			pass	pass	
Coverage histogram	-R	2 pass			pass	pass	
Cumulative genome coverage	R	1 pass			pass	pass	
Insert size histogram	Qualiman		orall state				
GC content distribution	Quaimap				action below for detailed	voou lito	
Picard	Guick summary	: Mean co	verage (min: 30)	Plot Sh	owing $\frac{3}{102}$ rows and $\frac{8}{14}$ colu	umns.	
Alignment Summary	Sample Name	mean	coverage val	avg gc	percentage aligned	mean coverage	median
Mark Duplicates		56.8	0 -	pass		pass	pass
WGS Coverage		26.7					
WGS Filtered Bases		20.7		pass	pass		
Samtools		33.3		pass	pass	pass	pass
Percent Mapped	Bcftools	stats					
Alignment metrics	Quick summary	of Bcftools	s-stats results.	See Bcftool	s section below for detail	ed results.	
Bcftools	Sopy table	III Configu	re Columns	Plot Sh	owing $^{3}/_{102}$ rows and $^{7}/_{14}$ colu	umns.	
Variant Substitution Types	Sample Name	numbe	r_of_records	number_	of_SNPs number_of_i	indels perc_snpa	s perc_
Variant Quality		pass		pass	pass	pass	pass
Indel Distribution	10.000	pass		pass	pass	pass	pass

fail

Figure 1. Aggregation and visualization of QC tools output and QuaC-Watch output using MultiQC at the project level. QuaC-Watch section shown here enables quick review of samples' QC results and helps identify samples that need further review. Additionally, similar MultiQC report is created at the single-sample level.

lds			
ups	bcftools	_stats	variant_per_(
	pass		pass
	pass		pass
	fail		fail
cores	per_b	ase_sequ	uence_cont
	fail		
	fail		
cover	age m	ean cov:	median cov
	P	ass	
	P	ass	
	P	ass	
ndels	tstv	heteroz	ygosity_ratio
	pass	pass	
	pass	pass	
	pass	pass	

Table 1. QC tools included in QuaC. Tools run directly by QuaC

 using BAM and VCF files as input are listed here. For practical reasons, QC for FASTQ files (fastqc, FastQ Screen) as well certain QC of BAM files (Picard-MarkDuplicates) are performed by the upstream alignment and small variant calling pipeline. However, QuaC can consume their output as well for the QC aggregation step.

Tool	Use
Qualimap	QCs alignment data in SAM/BAM files
Picard- CollectMultipleMetrics	Summarizes alignment metrics from a SAM/I file using several modules
Picard-CollectWgsMetrics	Collects metrics about coverage and performation of whole genome sequencing (WGS) experiment
mosdepth	Fast BAM/CRAM depth calculation
indexcov	Estimate coverage from whole-genome bam of cram index (Skipped in exome mode)
covviz	Identifies large, coverage-based anomalies (Skipped in exome mode)
bcftools stats	Stats variants in VCF
verifybamid	Estimates within-species (i.e., cross-sample) contamination
somalier	Estimation of sex, ancestry and relatedness

Table 2. Sample QC database. As part of the manual sample QC

 review, results are logged in to our internal sample QC database for various QC measures at multiple levels using controlled terms. Type 1 flags includes terms pass, acceptable, poor and fail. Type 2 flags includes terms pass, fail and not applicable. This database allows the downstream GS/ES data consumers to quickly identify problematic samples along with their potential cause for failure. This empowers them to quickly determine if such sample can be used for their intended data analyses.

Field	Explanation	Allowed values
Sample - Overall Status	Overall QC status considering results of all QC performed	Type 1 flags
FASTQ	Overall QC status considering results of all QC performed at FASTQ level	Type 1 flags
FASTQ Comment	Comments on QC at FASTQ level (e.g., small insert size, high adapter content, etc.)	Free text
BAM	Overall QC status considering results of all QC performed at BAM level	Type 1 flags
BAM Comment	Comments on QC at BAM level (e.g., low mean coverage, high duplication rate, etc.)	Free text
Other Species Contamination	Checks for sample contamination due to other species' genomic material	Type 1 flags
Human Cross- contamination	Checks for sample contamination due to other human's genomic material	Type 1 flags
Sex Check	Checks if predicted sex matches self-reported sex	Type 2 flags
Relatedness Check	Checks if predicted relatedness matches self-reported relatedness	Type 2 flags
Ancestry Check	Checks if predicted ancestry matches self-reported ancestry	Type 2 flags
Other Comments/Notes	Any other comments/notes concerning QC	Free text

Methods

- Pipeline component was developed using Snakemake, and QuaC-Watch and wrapper components were implemented using Python. It is a companion pipeline, which gets run after GS/ES samples are run through alignment and small variant calling, i.e., raw reads from FASTQs are aligned and then produced BAM and VCF files.
- QuaC isolates QC tool execution in a Conda environment within a Singularity container, as this setup provides the major advantage of reproducibility and portability.
- QuaC is run at a project level, and samples are provided as input in a pedigree file format (.ped), which allows to provide samples' relatedness and sex information to the pipeline.

eP350

QC type BAM QC BAM BAM QC BAM QC ance BAM QC BAM QC BAM QC VCF QC Within-species contamination Sex, ancestry and relatedness estimation