Methods

Multi-perspective quality control of Illumina exome sequencing data using QC3

Yan Guo a,⁎,1, Shilin Zhao a,1, Quanhu Sheng a, Fei Ye a, Jiang Li a, Brian Lehmann c, Jennifer Pietenpol c, David C. Samuels b, Yu Shyr a,⁎

a Vanderbilt Ingram Cancer Center, Center for Quantitative Sciences, Nashville, TN, USA
b Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN, USA
c Department of Biochemistry, Vanderbilt University, Nashville, TN 37027, USA

⁎ Corresponding authors.
E-mail addresses: yan.guo@Vanderbilt.edu (Y. Guo), shilin.zhao@vanderbilt.edu (S. Zhao), Quanhu.sheng@vanderbilt.edu (Q. Sheng), Fei.ye@vanderbilt.edu (F. Ye), riverlee2008@gmail.com (J. Li), brian.d.lehmann@vanderbilt.edu (B. Lehmann), j.pietenpol@vanderbilt.edu (J. Pietenpol), david.samuels@chgr.mc.vanderbilt.edu (D.C. Samuels), yu.shyr@Vanderbilt.edu (Y. Shyr).

⁎⁎ Equal contribution.

A R T I C L E   I N F O

Article history:
Received 14 October 2013
Accepted 25 March 2014
Available online 3 April 2014

Keywords:
Quality control
Exome sequencing
Raw data
Alignment
Variant call

A B S T R A C T

Advances in next-generation sequencing (NGS) technologies have greatly improved our ability to detect genomic variants for biomedical research. The advance in NGS technologies has also created significant challenges in bioinformatics. One of the major challenges is the quality control of sequencing data. There has been heavy focus on performing raw data quality control. In order to correctly interpret the quality of the DNA sequencing data, however, proper quality control should be conducted at all stages of DNA sequencing data analysis: raw data, alignment, and variant detection. We designed QC3, a quality control tool aimed at those three major stages of DNA sequencing. QC3 monitors quality control metrics at each stage of NGS data and provides unique and independent evaluations of the data quality from different perspectives. QC3 offers unique features such as detection of batch effect and cross contamination. QC3 and its source code are freely downloadable at https://github.com/slzhao/QC3.

1. Introduction

High throughput sequencing is the most effective way to screen for non-specific germline variants, somatic mutations, and structural variants. Some of the most popular sequencing paradigms in DNA sequencing are exome sequencing, whole genome sequencing, and target panel sequencing. While vastly informative, sequencing data poses significant bioinformatics challenges in areas such as data storage, computation time, and variant detection accuracy. One of the easily overlooked challenges associated with sequencing is quality control. Quality control for DNA sequencing data can be categorized into three stages: raw data, alignment, and variant calling [1]. One of the major misconceptions of DNA sequencing quality control is that quality control is only needed at the raw data stage. QC on raw sequencing data has been given more attention than QC on alignment and variant calling, and this is also reflected in the number of tools developed for conducting quality control on alignment and variant calling despite the fact that quality control is essential at all three stages of DNA sequencing data processing. Several notable alignment QC tools are SAMStat [8], and GATK’s QC measures [9]. SAMStat focuses on reads mapping statistics such as distribution of mismatches and insertions. When representing mapping statistics, SAMStat uses a pie chart which arbitrarily divides reads by map quality score. In our approaches, we use a distribution plot which does not arbitrarily stratify reads. GATK’s QC measures contain a series of independent scripts and tools for various purposes. The disadvantage would be running GATK multiple times in order to get complete QC results and GATK’s results are hardly interpretable for people with less experience. At the raw data stage, quality control serves as a quick screening for data with serious quality issues resulting from the sequencer. Tagging those data with questionable quality, Quality control at the alignment stage focuses on alignment quality, which is crucial for successful variant detection. Quality control at the variant calling stage is the last chance to identify samples with quality issues that slipped through quality control at earlier stages and to further reduce the number of false-positive variants.

We have previously detailed three-stage QC strategies in Briefings in Bioinformatics. Here, we present the actual implementation of these strategies in a single package: QC3, a quality control tool for all three stages of DNA sequencing data processing. QC3 provides both graphic and tabulated reports for quality control results. It also offers several unique features such as separation of good and bad quality reads (based on Illumina’s filter) and the detection of batch effects and cross
QC3 offers similar functionality as the existing tools which check base contamination of samples. QC3 accepts three types of data as input: FASTQ [10], Binary Alignment Map (BAM) [11], and Variant Calling Format (VCF), corresponding to the three stages of raw data, alignment, and variant detection respectively. The overview of the concept behind QC3 is shown in Fig. 1. QC3 is written with the combination of Perl and R and is freely available for public use. QC3 can be configured to run in series or parallel. We recommend running the complete QC in sequential order of the sequencing analysis. If needed, however, QC3 can perform the QC checks individually at each stage. QC3 can provide results with a reasonable time for processing. For Example, on a Linux machine with a single core CPU at 2.8 Ghz speed and 8 GB memory, QC3 can process a standard sequencing dataset (30–50 million pair-end reads) in around 20 min at each stage with some variation depending on file size.

2. Materials and methods

2.1. Raw data

Sequencing data analysis depends on raw data quality control for informative results, and there are many publicly available tools that suit this need. Investigators need to be aware of several considerations regarding these quality control parameters. The first is the Phred scale used to measure base quality. The standard scale is Phred +33 (ASCII 0 to 62), but older Illumina pipelines (before Casava 1.3) used Phred +64 (ASCII 59 to 126). Misidentifying the Phred scale [12,13] will result in errors. Correctly scaled data will have a median base quality score between 35 and 40 [14] for Illumina-generated data, and the shape of a base quality vs. cycle plot should be the same regardless of Phred scale. The second consideration is the nucleotide distribution across sequencing cycles. Ideally, four nucleotides should have a stable distribution across all reads, although there may be some fluctuation at the ends of reads or on the beginning cycles on some platforms. Nucleotide distribution is closely related to base quality, and an unstable nucleotide distribution will often accompany poor base quality. Both are good indicators of raw data quality. The third consideration is GC content. While the percentage of GC content varies across species and genomic regions, a significant deviation (over 10%) from the expected range of the sample may indicate contamination. For raw data quality control, QC3 offers similar functionality as the existing tools which check base quality distribution by cycle visualized by a Q-score vs. cycle plot, nucleotide distribution by cycle, and GC content.

An important feature that distinguishes QC3 from other QC tools is that QC3 performs data quality assessment by separating the low-quality reads from the high-quality reads. Low-quality reads are denoted with the letter “Y” in the 9th field in each read’s description row, which is contained in all FASTQ data generated by Illumina Casava 1.8+. Since the majority of current DNA sequencing data is paired-end, differences between read 1 and read 2 can provide information about read quality. Due to reasons such as phasing/pre-phasing, decreased signal to noise ratio, and template damage over many cycles of laser imaging, the second read sometimes has significantly worse quality than the first read. QC3 takes both types of reads into account and can detect discrepancies by calculating Pearson’s correlation and Euclidean distance between the base quality scores of the two reads. Significantly different base quality distribution between the read pair would indicate quality issue.

Quality control on raw sequencing data provides quick insight into sample quality and can potentially save a significant amount of time in later analysis by allowing early identification of questionable samples. Passing quality control at the raw data level, however, does not necessarily mean that a sample is free from quality issues. Nor does questionable quality within a sample mean that the sample is necessarily useless for subsequent analysis. For example, a portion of reads in a sample can be bad, causing the sample to fail the raw data quality control checks, but after removing those bad reads, a sufficient number of good quality reads may still be present to allow further meaningful analysis to be carried out. Raw data quality control is necessary and informative, but one cannot determine the sample quality solely based on the raw data quality control results.

2.2. Alignment

Alignment is a non-optional step for any re-sequencing analysis. Quality control on the alignment provides additional insights into sample quality and can help identify bad samples that slip through the raw data quality control checks. Despite this, quality control of the alignment stage is not formally included in many sequencing data analysis pipelines and is not performed on a regular basis. For exome sequencing, one of the most important QC parameter is the capture efficiency,
defined as the number of reads aligned to capture regions divided by the total number of reads after some quality filter. Capture efficiency is highly dependent on the capture kit. For exome sequencing, there are three major exome sequencing capture kits on the market: Illumina TrueSeq, Agilent SureSelect, and NimbleGen SeqCap EZ. The capture regions for the exome capture kits range from 37.6 to 62.1 million base pairs. Other capture techniques available include array-based and selector probe-based methods. Previous studies have shown that capture efficiencies between 40 and 70% are typical for exome sequencing [14–17]. Lower capture efficiencies can indicate low complexity in the target library, sub-optimal probe hybridization conditions, or low stringency washes post-capture. QC3 evaluates capture efficiency if the capture kit’s BED file is provided. Otherwise, QC3 uses the standard exome annotation file from RefSeq to define the exome regions as capture regions.

Furthermore, it has been shown that reads aligned in intron and intergenic regions can generate high-quality data [17], and mitochondria genome sequence can be extracted from exome sequencing data [18]. Thus, we designed QC3 to separate reads into several different categories (on target, intronic, intergenic, mitochondrial, and unmapped) and to compute quality control parameters such as the median and mean mapping quality, insert size, and depth for each category. Median depth is not calculated by default due to the long computation time required, but the user can choose to run the calculation. As informative as quality control on alignment can be, contaminated samples might not be detectable in this stage.

2.3. SNP detection

For the majority of exome sequencing studies, detecting SNPs is one of the pivotal steps leading toward the final conclusion of the study. Quality control on SNP calls will not only help identify bad samples that have passed through the raw data and alignment quality control checks but will also minimize the rate of false positive SNP calls. QC3 uses several important criteria for evaluating SNP calls, including the transition/transversion (Ti/Tv) ratio and the heterozygosity to non-reference homozygosity ratio (het/non-ref-hom). The Ti/Tv ratio has been used by multiple studies [18,19] as a quality control parameter for assessing the overall SNP quality. The Ti/Tv ratio is computed as the number of transition SNPs divided by the number of transversion SNPs. Even though the number of possible transversions is twice as many as the number of possible transitions, leading to a Ti/Tv ratio of 0.5 if mutations occurred at equal rates, the actual Ti/Tv ratio differs by genomic regions. The Ti/Tv ratio is around 3.0 for SNPs inside exons and about 2.0 elsewhere [14,20], and the ratio also differs between synonymous and non-synonymous SNPs [21]. Because the target regions of exome capture kits often cover more than just exons, the Ti/Tv ratio for SNPs inside these target regions is expected to lie between 2.0 and 3.0 with the

Fig. 2. A. Distribution of read count by lane identifies lane 7 as a clear outlier. Lane 7 contains sample 37, which was later determined to have failed sequencing. B. Base quality by cycle plot for sample 37 fails to detect sample 37 as an outlier. C. Nucleotide distribution by cycle plot for sample 37 does not detect sample 37 as an outlier. D. GC content by lane shows that samples on lane 8 have a slightly lower GC content, causing us to flag these samples.
value depending on the fraction of exons inside target regions. Ti/Tv ratios in exome sequencing below the 2–3 range may be cause for concern. The het/non-ref-hom ratio is another good quality control parameter for DNA sequencing data. We have mathematically proved that for whole genome sequencing data het/non-ref-hom ratio is 2.0 based on the Hardy–Weinberg equilibrium [1]. For exome regions het/non-ref-hom would be significantly lower.

QC3 also checks the number of non-synonymous SNPs. In 2009, Bamshad et al. [22] showed that about 200 novel nonsynonymous SNPs should be expected per person through exome sequencing when compared to dbSNP 131. The most updated dbSNP version 137 contains 72,952,578 additional SNPs, a 63% increase compared to version 131. Thus, the number of novel nonsynonymous SNPs should be less than what Bamshad et al. reported in 2009. An excessive number of novel nonsynonymous SNPs detected would raise a quality control concern. QC3 performs annotation against dbSNP 137 and reports novel non-synonymous SNPs by sample using the ANNOVAR tool [23].

QC3 is capable of detecting cross-contamination of samples, which can slip through the raw data and alignment quality control checks without drawing any attention. Cross-contamination happens when the DNA of different samples are accidentally mixed. QC3 detects such problematic samples by computing the genotype consistencies between all possible sample pairs. Three different types of genotype consistency are computed: 1) heterozygous consistency using sample 1 in the pair as denominator; 2) heterozygous consistency using sample 2 in the pair as denominator, and 3) overall consistency including both heterozygous and homozygous genotypes. The genotype consistency checking is similar to the idea of computing identity by descent (IBD) in a genetic study. Excessive heterozygous consistency between two random samples indicates cross contamination, or unknown relatedness between the samples. Because majority of the human genome is homozygous, the overall consistency should be much higher than heterozygous consistency. Low heterozygous or overall consistency between known related samples can indicate pedigree error or other contamination.

Additionally QC3 is able to estimate gender. Gender is estimated from two perspectives: read count in Y chromosome and proportion of heterozygous SNPs in X chromosome. Females do not have the Y chromosome. Thus, for a female, the reads aligned to Y chromosome should be significantly less than that of a male. Some reads may still align to Y chromosome due to homologous regions. Males have two X chromosomes. Thus, all SNPs on X chromosome should be homozygous. Excessive heterozygous SNPs on the X chromosome indicate the sample is from a female host. By examining both read count on Y chromosome and the proportion of heterozygous SNPs on X chromosome, we can make a reasonably accurate inference on gender.

2.4. Batch effect detection

Batch effects are systematic technical variations in data caused by processing data in batches. It can cause a complication during data analysis especially for a large dataset. The Cancer Genome Atlas (TCGA), one of the largest public available datasets also documented the evidence of batch effect. The most common sequencing failures often occur non-randomly, by lane, flow cell, run, or machine. QC3...
recognizes and records the machine name, run ID, flow cell ID, and lane ID of an experiment from either the FASTQ file or BAM files. Based on this information, QC3 determines whether a batch effect exists using a nonparametric Kruskal–Wallis [24] test and a Fligner–Killeen test of homogeneity of variances [25] on the QC parameters (read count, mapping quality, base quality, and capture efficiency).

3. Results

To demonstrate the effectiveness of QC3, we performed three-stage quality control using QC3 on a breast cancer exome sequencing study. This study contains a total of 30 breast tumor samples. The exomes were captured using Illumina’s TrueSeq capture kit and the libraries were sequenced on a HiSeq 2000 platform. The reads are 75 base-pairs long and with a pair-end.

We first performed the raw data quality control on the raw data of the 30 samples using QC3. Immediately, we noticed that sample 37 had only 14,353 reads, which clearly indicated failed sequencing. Interestingly, sample 37 was not picked up on the base sequence quality plot or nucleotide distribution plot, which most raw data quality control programs draw (Figs. 2B,C). By checking the batch effect, the low read count of sample 37 also skewed the read count distribution of lane 7, which was captured by the Fligner test (p-value < 0.001) (Fig. 2A TotalReadsByLane). Other interesting findings from the batch effect test were relatively lower GC content and base quality for samples on lane 8 (Fig. 2D). These results prompted us to re-sequence sample 37 and flag samples run on lane 8. For demonstration purposes, we included the failed sample 37 for the later quality control steps. The tabular report of raw data quality control can be viewed in Table S1.

After raw data quality control, we performed alignment using BWA [26] and performed alignment quality control on the aligned BAM files. Unsurprisingly, sample 37 was picked out as the obvious outlier (Fig. 3A). In addition, sample 46 showed lower mean mapping quality, which caused us to flag it (Fig. 3B). Furthermore, we did not detect any abnormal behavior on alignment statistics for the flagged samples on lane 8. The complete tabular report can be found in Table S2.

SNPs were called using GATK’s Unified Genotyper [9]. The VCF files were quality controlled by QC3. The Ti/Tv ratio and heterozygous/non-reference homozygous ratio of all samples appear normal even for

---

**Fig. 4.** A. Heatmap displays pair-wise sample heterozygous genotype consistency using the first sample in the pair as a denominator. Pairwise comparison identifies possible cross-contamination between samples 10 and 14, and samples 39 and 40. B. Heatmap matrix displays pair-wise sample heterozygous genotype consistency using the second sample in the pair as a denominator. C. Overall pair wise consistency which is computed as the number of consistent genotypes divided by the number of total shared positions with genotypes called for both samples.
sample 37 (Table S3). This shows that even though the number of reads is too few to be considered a successful sequencing, the few reads that did get sequenced produced a few hundred reliable SNPs. Through contamination analysis, we found possible cross-contamination between samples 10 and 14, and samples 39 and 40 (Figs. 4A,B). The heterozygous consistency rate of 0.91 and 0.99 were observed between these two pairs of samples respectively. The normal heterozygous consistency rate between two unrelated samples should be around 0.2 to 0.3. Also sample 43 showed high heterozygous consistencies with multiple samples which also indicate possible contamination. No outliers were observed for the samples run on lane 8, despite observing slight abnormalities for these samples during raw data quality control. Therefore we conclude that these samples are useable. The complete tabular report of contamination can be found in Table S4. QC3 organized the figures and tables into a HTML file for better readability. The figures and tables presented here only represent a partial output of QC3’s results.

A second example was run using exome sequencing data from 11 TCGA breast cancer samples. TCGA data has been through rigorous quality control before releasing to the public. Thus, it would generally not have any quality issue. The samples were chosen randomly without any bias. Out of the 11 samples, 10 are from tumor tissue, and the 11th was a blood sample from the same individual already presented in the previous 10. The reason to pick the 11th sample from the same individual was to artificially create a situation where two samples have a very similar heterozygous SNP consistency which resembles the scenario of cross contamination. As expected, raw data and alignment QC on this dataset showed great quality. However by running VCF QC, we successfully identified that two samples are closely related through their high heterozygous SNP consistency rate. This example once again demonstrates that QC on raw data and alignment do not identify cross contamination problems. The complete QC report of raw data, alignment, VCF and information on the TCGA sample’s barcode can be viewed at [https://github.com/slzhao/QC3](https://github.com/slzhao/QC3).

4. Discussion

Illumina’s sequencing platform has dominated the sequencing market for the last few years with no sign of diminishing. Thus, we developed our tool based on the Illumina sequencing platform. However, the general concepts discussed here are applicable across a range of sequencing platforms, with appropriate modifications where necessary. QC3 is designed with a large sample size in mind. Although it works with single samples in general, some of its features such as detection of contamination and batch effect are only appropriate for large sample size studies.

As we have demonstrated, QC3 performs quality control at three different stages and from different perspectives. It is essential that we do not make hasty decisions based on the quality control results from just a single perspective. Evaluating high throughput data quality from the perspectives of raw data, alignment, and variant calling allows researchers to identify problematic samples with greater confidence.

Even with the most rigorous quality control protocol, there are still certain false positive results that can evade our quality control efforts. For high impact studies, the use of additional methods such as RT-PCR, Sequenom, or Sanger sequencing to validate the findings independently from NGS sequencing is highly recommended.

5. Availability

The source code, instruction manual and examples used in the manuscript of QC3 can be accessed at [https://github.com/slzhao/QC3](https://github.com/slzhao/QC3).

Competing interest statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Acknowledgment

This study was supported by CCSG (P30 CA068485). We would also like to thank Margot Bjoring for editorial support.

References