BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 6. 879–889
Advance Access published on 24 September 2013
doi:10.1093/bib/bbt069

Three-stage quality control strategies for DNA re-sequencing data

Yan Guo, Fei Ye, Quanghu Sheng, Travis Clark and David C. Samuels

Submitted: 13th June 2013; Received (in revised form): 21st August 2013

Abstract

Advances in next-generation sequencing (NGS) technologies have greatly improved our ability to detect genomic variants for biomedical research. In particular, NGS technologies have been recently applied with great success to the discovery of mutations associated with the growth of various tumours and in rare Mendelian diseases. The advance in NGS technologies has also created significant challenges in bioinformatics. One of the major challenges is quality control of the sequencing data. In this review, we discuss the proper quality control procedures and parameters for Illumina technology–based human DNA re-sequencing at three different stages of sequencing: raw data, alignment and variant calling. Monitoring quality control metrics at each of the three stages of NGS data provides unique and independent evaluations of data quality from differing perspectives. Properly conducting quality control protocols at all three stages and correctly interpreting the quality control results are crucial to ensure a successful and meaningful study.

Keywords: *sequencing; quality control; FASTQ; alignment; variant calling*

BACKGROUND

High-throughput sequencing is the most effective way to screen for non-specific germ line variants, somatic mutations and structural variants. Some of the most popular sequencing paradigms in DNA sequencing are whole-genome sequencing, exome sequencing and target panel sequencing. While vastly informative, sequencing data poses significant bioinformatics challenges in various areas such as data storage, computation time and variant detection accuracy. One of the major challenges associated with sequencing data that is sometimes easily overlooked is monitoring quality control metrics over all stages of the data processing pipeline. Quality control for DNA sequencing data can be separated into three stages: raw data, alignment and variant calling. A common misconception of DNA sequencing quality control is that quality control is only needed at one or two of these stages. There is usually a focus on quality control at the raw data stage rather than the alignment and the variant calling. However, quality control is essential to all three stages: At the raw data stage, quality control serves as a quick screening for excluding data with serious quality issues and flagging data with questionable quality. Quality control at the alignment stage focuses on the alignment quality, which is crucial for successful variant detection. Quality control on variant calling is the last chance to identify samples with quality issues that are not detected at earlier stages and to further reduce false-positive variants.

In this article we will discuss quality control strategies at each of the three stages, focusing on human DNA re-sequencing data. Although DNA re-sequencing technologies have been used on other species such as viruses, bacteria and plants, because of the lack of precise annotation and exome extraction kits, some of the strategies described in this article may be difficult to apply to these species. Also, because Illumina's sequencing platform has

Corresponding author. Yan Guo. 2220 Pierce Ave, 549 Preston Research Building, Nashville TN 37232, US. E-mail: yan.guo@vanderbilt.edu

Yan Guo is an assistant professor at Department of Cancer Biology, Vanderbilt University. He is the Technical Director of Bioinformatics for Vanderbilt Technologies for Advanced Genomics Analysis and Research Design.

Fei Ye is an assistant professor at Department of Biostatistics, Vanderbilt University.

Quanghu Sheng is a post doc fellow at Department of Cancer Biology, Vanderbilt University.

Travis Clark is the technical director of Next-Generation Sequencing at Vanderbilt Technologies for Advanced Genomics.

David Samuel is an associate professor at Department of Molecular Physiology & Biophysics, Vanderbilt University.

dominated the sequencing market for the past few years with no signs of diminishing, we will focus our review on the Illumina sequencing platform; the general concepts discussed here, however, are applicable across a range of sequencing platforms, with appropriate modifications where necessary.

QUALITY CONTROL OF THE RAW DATA

Raw data quality control should be the initial step of data analysis for any successful study. There are several tools that are publically available for conducting quality control on raw FASTQ files. One of the pioneers in raw sequencing data quality control was the FASTX-Toolkit, which is a collection of Linux command line tools for processing FASTQ files. This tool is capable of checking base quality and nucleotide distribution. A more advanced tool dealing with raw data quality control is the FastQC package developed by the Babraham Institute bioinformatics group. FastQC offers some additional quality control parameters that are not included in the FASTX-Toolkit, including the average base quality score per read, the GC content distribution and identification of the most duplicated reads. More importantly, FastQC can use aligned BAM [1] files instead of FASTQ files to assess the quality control of raw data. Other similar raw data quality control tools are next-generation sequencing (NGS) Quality Control (QC) Toolkit [2], RRINSEQ [3] and QC-Chain [4]. FastQ Screen is another tool developed by the Babraham Institute's bioinformatics group, which can be used to screen for cross-species contamination using FASTQ files.

The most important parameters to check for raw sequencing data quality are the base quality, the nucleotide distribution, GC content distribution and the duplication rate. A common way to visualize base quality is to draw a base Q-score versus cycle plot. Sequencing data generated on Illumina platforms tend to observe a median base quality score between 35 and 40 [5] in the Phred scale [6, 7]. The older Illumina pipeline (Before Casava 1.3) used Phred +64 (ASCII 59–126) instead of the standard Phred +33 (ASCII 0–62). Investigators need to be aware of the exact scale of the Phred score used when choosing a quality control tool. For example, the FASTX-Toolkit will give an error if Phred +64 FASTQ data are input to that tool. Although the Phred scale maybe different, the shape of the figure (Base Quality (BQ) versus Cycle) should remain exactly the same. Outliers in the data can be identified graphically, regardless of the scale.

For the older Illumina sequencing platform GA II, the base quality usually starts out high then gradually drops as the cycle increases (Supplementary Figure S1a). This is due to factors that impact the accuracy of the base-calling algorithm on the cluster over time, such as phasing/pre-phasing, decreased signal to noise ratio and template damage over many cycles of laser imaging. For the newer HiSeq 2000 and 2500 systems, a similar pattern of decreasing base quality at the end of reads can be observed. However, owing to changes in Illumina's quality score algorithm, the base quality for the first 10–15 cycles are relatively lower compared with the middle section of the read (Supplementary Figure S1b). The common way of dealing with the low-quality bases at the end of the read is by trimming [8], i.e. removing bases within a certain distance from the read ends. For all Illumina sequencing platforms, the median base quality score should stay >30 across the length of the reads. Large variations in base quality scores (Supplementary Figure S1c) usually indicate that many low-quality reads were sequenced from low-quality DNA samples, and a sudden drop in quality (Supplementary Figure S1d) can indicate adaptor contamination or fluidics problems during the run. Early HiSeq 2000 instruments often had solenoid clogs that impacted one or more cycles during a run. For paired-end reads, it is common to observe higher quality in the first end of the read than the second end owing to the amount of time the template was on the instrument and increasing laser exposure over time.

The nucleotide distribution across cycles is another useful quality control parameter for whole genomes and exomes but not amplicons or RNA-seq samples. For a perfect sequencing run, the distribution of the four nucleotides (A T C G) across all reads should remain relatively stable (Supplementary Figure S2a), except for minor fluctuations at the end of the read. The HiSeq 2000 also shows some fluctuation for the beginning 10–15 cycles where cluster identification and assignment is being performed by the RTA software (Supplementary Figure S2b). The nucleotide distribution is closely associated with base quality, and they can both be used to measure the quality of the raw data. Data with bad base quality are usually also reflected in the nucleotide distribution plot. Using the same data from

Supplementary Figure S1a–d, an example of nucleotide distribution from a bad quality DNA sample can be seen in Supplementary Figure S2c, and an example of contamination can be seen in Supplementary Figure S2d.

The total percentage of GC content sequenced can also be used as a quality control parameter. The percentage of GC in the genome varies across species and across the regions of each genome. For exome regions, the GC content is about 49–51%, while for whole-genome sequencing, the GC content is around 38–39%, and for *Saccharomyces cerevisiae* and *Mycobacterium tuberculosis*, the GC content is around 38–42%. Abnormal GC content percentage, say, more than 10% deviation from normal range, can indicate contamination.

There are many sources that could introduce errors during the sequencing process. One of the most common sources for quality concern is the DNA library quality, which can also be directly reflected by the number of reads sequenced. For exome sequencing, the most popular platform is Illumina's HiSeq 2000, which is capable of generating 37.5 Gb of data (~375 M reads) per lane. A common way to conduct exome sequencing cost efficiently is to multiplex and pool three to four samples on one HiSeq 2000 lane. Even though the pool is constructed with a target of uniform molarity, each sample's DNA contribution to the pool is never equal. The number of reads sequenced per sample is highly dependent on the quality of a sample's DNA quantitation or DNA library quality. It has been reported that the difference in the number of reads between multiplexed samples on a single lane can be as large as 2- to 4-fold [5, 9]. Variability in the DNA library amount in a pool is owing to quantitation and mixing, not quality of the input DNA to the library prep. The yield per sample is the amount of sequencing data generated per sample. The yield per sample for multiplexed samples is most often reported as the pass filter yield after demultiplexing. A high variation in the yield per sample can indicate either improper pooling or problems with the sequencing of the pooled samples. For example, during a breast cancer exome sequencing study, 24 samples were sequenced. By drawing the read count by lane (Supplementary Figure S3), it is clear that one sample had a much lower read count than the others did.

Quality control on the raw sequencing data provides a quick insight into the sample quality and can save a significant amount of time in later analysis by allowing early identification of bad samples. However, passing quality control at the raw data level does not necessarily mean that a sample will pass alignment quality control checks. On the other hand, a sample with some raw data quality control problems might still be salvageable for later analysis. Sometimes, a portion of reads in a sample can be bad, which will cause it 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 analysis to be carried out. Raw data quality control is necessary and informative, but one cannot determine the sample quality purely based on the raw data quality control results.

ALIGNMENT QUALITY CONTROL

Alignment is a non-optional step for any re-sequencing analysis. It provides additional insights into sample quality and can help identify bad samples that pass the raw data quality control checks. However, alignment for quality control is not performed on a regular basis. Different alignment quality control parameters should be collected for exome sequencing and whole-genome sequencing. For exome sequencing, there are three major exome sequencing capture kits in 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 including array based and selector-probe based methods are also available. The capture efficiency varies by capture method. Capture efficiency is the most important quality control parameter for exome sequencing or other targeted sequencing. Previous studies have shown that capture efficiencies between 40 and 70% are typical for exome sequencing [5, 10–12]. To further investigate this, we performed a capture efficiency analysis on >600 breast cancer samples (with the majority from TCGA breast cancer cohort [13] and a few from a previous study [5]). Figure 1 shows the distribution of the source of the captured DNA for three popular exome-sequencing platforms: Agilent SureSelect, Illumina TrueSeq and NimbleGen SeqCap EZ capture kits. After filtering out low-quality reads, only 50–75% of the captured reads were from the target regions. Based on RefSeq release 57, ~2% of the reads mapped to untargeted exon regions while 12–20% mapped to introns and 10–25% mapped to intergenic regions. A small sliver of the reads, 0.004–

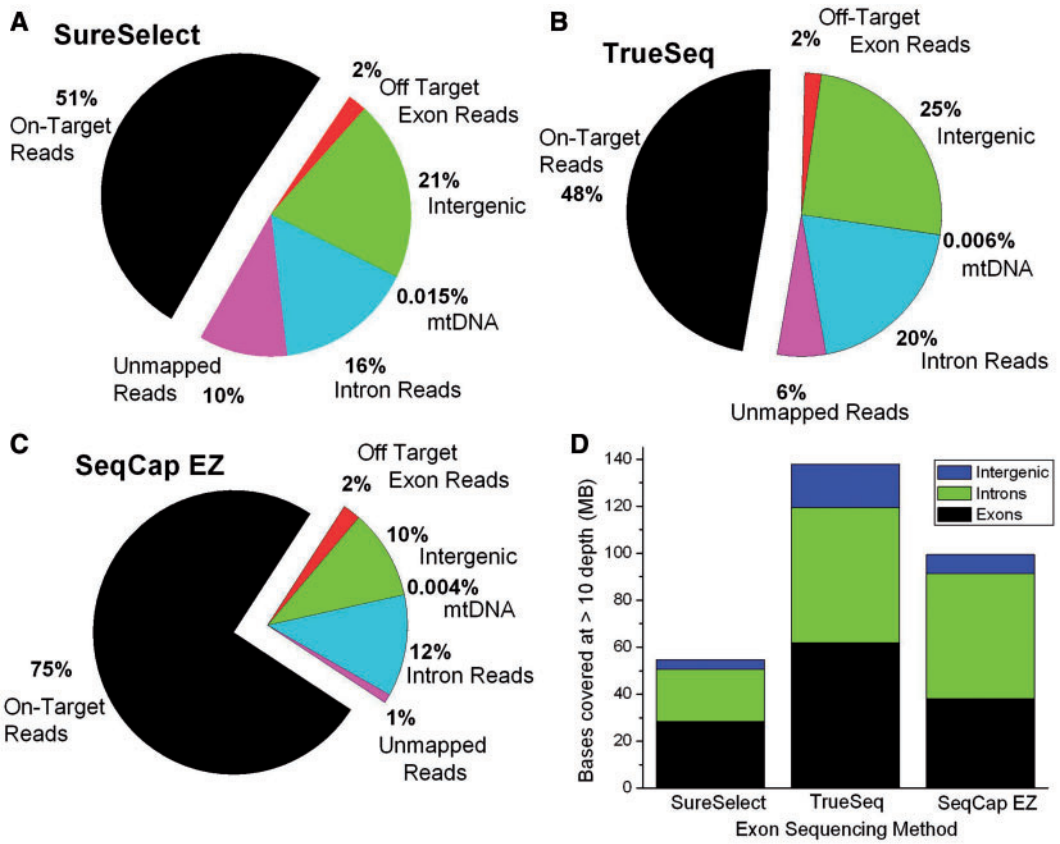

Figure 1: The percentages of reads assigned to different categories for (**A**) SureSelect (v2), (**B**) TrueSeq and (**C**) the SeqCap EZ methods of exome sequencing. In all cases, the largest category of reads consists of the targeted genomic regions, but a large fraction of the reads are off target. The categories shown are the reads that map to exons that were not part of the target set, intergenic regions, mtDNA, introns and finally reads that do not map to any part of the human reference sequence. (**D**) The total number of bases covered at >10 depth that map to exons (both targeted and non-targeted exons), introns and intergenic regions for three methods of exome sequencing. These numbers should be compared with the full human genome size of approximately 3 billion base pairs.

0.015%, mapped to the mitochondrial genome. Finally, an appreciable portion of the reads (1–10%) did not map to the human reference genome at all. Lower capture efficiencies indicate low complexity in the target library, suboptimal probe hybridization conditions or low stringency washes after capture. Currently, there are two tools available to perform alignment quality control: Picard and QPLOT (Table S1).

Additional quality control parameters to check for exome sequencing alignment are the median depth, mapping quality (the 5th field of a BAM file), insert size (the 9th field of a BAM file) and the number of discordantly mapped pairs.

The most important quality control parameter for whole-genome sequencing is the average or median depth and the percentage of the genome covered by the sequencing at that depth. For example, the Illumina service lab promises whole-genome sequencing with an average depth of 30 across 98% of the genome. However, average depth is a misleading term because it can be skewed easily by the high-depth regions. During exome sequencing, there is a phenomenon called unspecific binding where certain regions of the genome have much higher than usual depth. A previous study has shown that exome sequencing data from multiple capture technologies produced an implausible 500×–1000× depth in such regions compared with an average of 30× depth in most of the capture regions [5]. Such regions can skew the average depth statistic. The commonly used Genome Analysis Tool Kit (GATK) [14] developed by Broad Institute also suggests excluding such regions for variant calling. Median depth would be a more robust statistic, though this is rarely used in practice owing to the long computation time required.

QUALITY CONTROL ON VARIANT CALLING

For the majority of exome sequencing studies, detecting Single Nucleotide Polymorphisms (SNP) is one of the pivotal steps leading towards the final conclusion of the study. Quality control on SNP calls will not only help identify bad samples that have slipped through raw data and alignment quality control checks but will also minimize the rate of false-positive SNP calls. There are several situations such as cross-contamination and mislabelling where a bad sample can pass through the raw data and alignment quality control checks. Cross-contamination happens when the DNA of different samples are accidentally mixed. Mislabelling happens when samples are switched owing to human error. Both scenarios produce DNA that do not represent the original intended sample and can produce high-quality raw data and alignment. Identities by Descent or simple genotype consistency are useful statistics to detect bad samples caused by cross-contamination. Mislabelling is impossible to catch, unless additional samples from the same pedigree are sequenced.

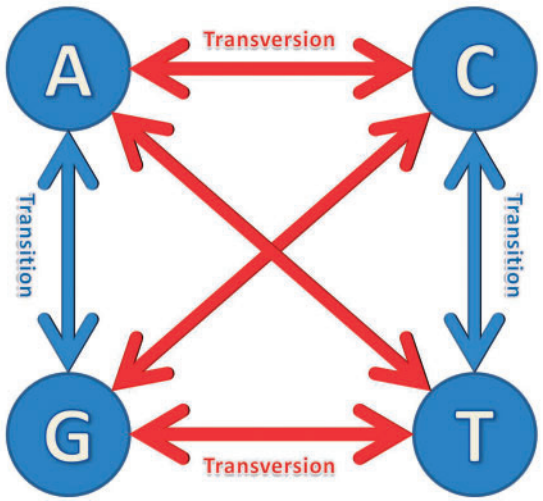

Figure 2: The Ti/Tv ratio is computed as the number of transition SNPs divided by the number of transversion SNPs. Transitions involve interchanges of nucleotides of similar shapes: two-ring purines (A←→G) or one-ring pyrimidines (C←→T). Transversions involve interchanges of one-ring and two-ring structures (A←→C, A←→T, G←→T, G←→C). 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 transition/transversion (Ti/Tv) ratio (Figure 2) has been used by multiple studies [5, 15, 16] as a quality control parameter for checking the overall SNP quality. The Ti/Tv ratio is computed as the number of transition SNPs divided by the number of transversion SNPs. Transitions involve interchanges of nucleotides of similar shapes: two-ring purines (A←→G) or one-ring pyrimidines (C←→T). Transversions involve interchanges of one-ring and two-ring structures (A←→C, A←→T, G←→T, G←→C). 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. For human genome data, the Ti/Tv ratio is around 3.0 for SNPs inside exons and about 2.0 elsewhere [17], and the ratio also differs between synonymous and non-synonymous SNPs [18]. 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 value depending on the fraction of exons inside target regions. Ti/Tv ratios in exome sequencing below the two to three range may be cause for concern. When computing the Ti/Tv ratio it is important to be aware of the plausible Ti/Tv ratio range and any potential bias that might skew it. For example, mitochondrial DNA (mtDNA) has been reported to have a much stronger bias towards transitions over transversions compared with nuclear genes [19, 20]. Thus, it is a good practice to exclude mtDNA when computing the Ti/Tv ratio. Also, when sequencing non-human samples such as plants and bacteria, an investigator should look up any previous specific knowledge of the Ti/Tv ratio for that species. For example, it has been reported that the Ti/Tv ratio is 0.62 for yeast [21], and 1.5 for maize [22]. If a lower than expected Ti/Tv ratio is observed, we suggest that the investigator apply more stringent criteria on filters such as depth and genotype quality and then compute the Ti/Tv ratio again to see if these increased quality control (QC) criteria have caused any improvement in this ratio.

The number of novel non-synonymous SNPs can also be a good indicator of the false-positive rate. Bamshad *et al.* [23] (in 2009) showed that about 200 novel non-synonymous SNPs should be expected per person through exome sequencing and that a higher number would likely indicate a high false-positive rate. The novel SNPs identified

by Bamshad *et al.* were identified against Single Nucleotide Polymorphism database (dbSNP) 131. Currently, dbSNP has been updated to version 137, with 72 952 578 more SNPs compared with version 131 (a 63% increase since 2009). Thus, the number of novel non-synonymous SNPs should be less than what Bamshad *et al.* reported in 2009. This variability in the definition of 'novel' over time is a problem with using this measure as a quality control check. However, a significantly large number (>200) of novel non-synonymous SNPs would indicate a high false-positive rate. A good practice is to annotate called SNPs using tools such as ANNOVAR [24] and then determine if there are too many novel non-synonymous SNPs against the latest version of dbSNP.

Genotype consistency between SNP chip data and exome sequencing data has been used by multiple studies as a quality control parameter when genotyping data from both SNP chips and exome sequencing are available [5, 15]. The most popular DNA sequencing processing pipeline, GATK [14], implemented variant quality score recalibration based on SNP chip data to improve SNP quality called from exome sequencing. The overall consistency is computed as the number of discordant SNPs between the SNP chip and exome sequencing divided by the total overlapping SNPs between the SNP chip and exome sequencing. Overall consistency is misleading because it is easily inflated by the high number of homozygous loci in the human genome. In comparison, the heterozygous consistency rate is more informative and should be used instead of an overall consistency rate. The heterozygous consistency rate is computed as the number of heterozygous discordant SNPs between an SNP chip and exome sequencing divided by the total number of heterozygous SNPs overlapped between the SNP chip and exome sequencing. An example of the heterozygous consistency rate computation at different depths and genotype quality cut-offs is given in Supplementary Table S2.

A potential problem associated with using SNP chip consistency as a quality control parameter is the selection bias for the SNPs contained on the SNP chip. Guo *et al.* [16] pointed out that the SNPs present in both the SNP chip and the exome sequencing data are only a small fraction of the SNPs identified by exome sequencing, and the SNP chip's selection criteria selected SNPs that are more easily sequenced by exome sequencing technology. The overlapping subset of SNPs will therefore have better quality than the rest of the SNPs identified by exome sequencing, and thus the consistency rate for this special subset of SNPs might not represent the overall quality of SNPs identified by exome sequencing.

The heterozygosity to non-reference homozygosity ratio is another good quality control parameter

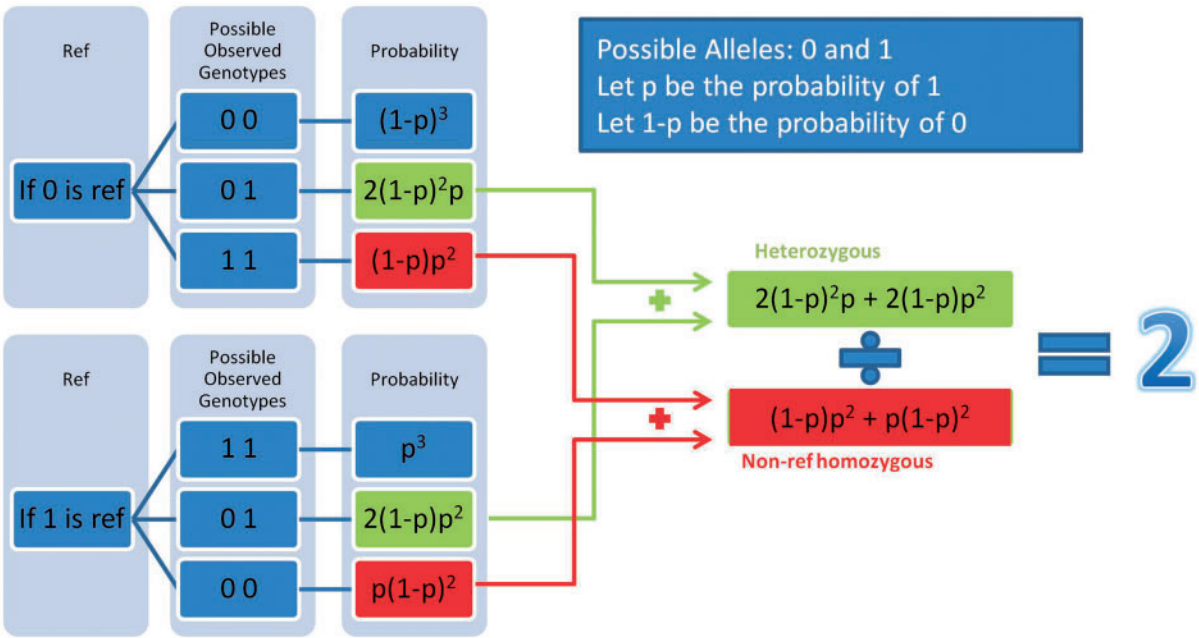

Figure 3: Proof of the principle that the heterozygosity to non-reference homozygosity ratio equals 2 for whole-genome sequencing data.

for DNA sequencing data. For whole-genome sequencing data, this ratio is 2.0 for variants in Hardy–Weinberg equilibrium. We provide a derivation for this value in Figure 3.

The aforementioned quality control parameters such as the Ti/Tv ratio, SNP chip consistency, heterozygosity rate and novel SNP number evaluate overall the SNP quality at a per sample level. There are numerous quality parameters that can evaluate SNP quality at a per SNP level. The most obvious parameter to use is the depth because higher depth gives more statistical confidence to the SNP call. Base quality and mapping quality filters can also be used to prevent bad reads from contributing to SNP calls. Base quality scores are computed by Illumina's sequencing platform, as aligners do not change the base quality scores. However, mapping quality scores are computed differently by each aligner. For example, BWA [25] computes mapping quality (range: 0 to 70), which is intended to reflect the actual quality of the alignment. On the other hand, Bowtie 2 [26] uses mapping quality (range: 1–255) to denote the uniqueness of the alignment, where 1 means the largest number of possible locations for alignments reported, and 255 means only 1 possible location for the alignment reported. Bowtie 2 reports the real mapping quality in the optional Alignment Score field in the BAM file. It is important to understand the actual implementation of mapping quality used by each aligner. The commonly used base quality score threshold for BWA is 20, while the mapping quality score threshold varies by aligner. It is recommended to draw the distribution of mapping quality scores and examine this distribution for outliers in cases where no previously suggested cut-off value for the mapping quality score is available.

Based on the Hidden Markov Model, the author of BWA and Samtools, Heng Li, introduced the Base Alignment Quality (BAQ) score [27], which reduces the false-positive SNP calls by decreasing the base quality scores for bases around insertion and deletion events in the sequence because indels often lead to alignment artefacts. This BAQ scoring system has been applied by the 1000 Genomes Project and implemented into Samtools as a default parameter. In a separate study, Guo *et al*. have shown that if BAQ and GATK's local realignment are used consecutively, instead of reducing the false-positive SNP calls additional false-positive SNPs can be introduced [16]. This is because both BAQ and GATK's local realignment aim to correct false SNPs caused by insertions and deletions, and so applying both at the same time will cause an over-correction. In many cases, people unknowingly apply both BAQ and GATK's local realignment because Samtools' BAQ option is turned on by default.

Once the variants are called, GATK recommends a set of filters in their best practice protocol. Based on the latest version of the protocols, the filters for SNPs include Quality by Depth (QD) <2.0, RMS Mapping Quality <40.0, HaplotypeScore >13.0, MQRankSum <−12.5 and ReadPosRankSum <−8.0. The filters for indels include QD < 2.0, ReadPosRankSum <−20.0, InbreedingCoeff <−0.8 and Fisher Score >200.0. These filters have been used in many research studies [28–30]. The detailed description of each filter is given in Table S3. All of these filters are limited to GATK's variant caller result and are not applicable to results generated by other variant callers.

There are also other lesser known quality control parameters for checking SNP quality such as strand bias, allele balance and cycle bias. Strand bias is the phenomenon when the genotype inferred from the positive strand and negative strand are significantly different, with one homozygous and the other heterozygous. In a study by Guo *et al*. [16], the authors showed that extreme strand bias creates false-positive SNPs and compared several different ways to compute strand bias. Strand bias has been used as a filter in a study related to mitochondria mutation and radiation [19], and a strand bias score is now computed in several variant callers such as Samtools [1], VarScan [31], SomaticSniper [32], MitoSeek [33] and MuTect [34].

For sequencing data, the reads at heterozygous SNPs should have an allele balance of 50%, meaning that 50% of the reads should support the reference allele while the other 50% of the reads should support the alternative allele. The percentage of alternative alleles observed is an important factor for all SNP callers, and it is often affected by reference allele preferential bias. Reference allele preferential bias is a phenomenon during alignment where there is preference towards the reference allele caused by alignment algorithms that penalize a mismatch from the reference. Degner *et al*. described such a bias in RNAseq data [35], and Guo *et al*. also described this in exome sequencing data [36]. The bias is anywhere from 1 to 5%. It is good practice to account for this bias when calling SNPs based on alternative allele percentage by adjusting the SNP detection

threshold slightly in favour of the non-reference allele.

SNP density is also an informative parameter. The frequency of observed SNPs within a fixed range of genomic regions should be reasonable. A high SNP frequency in a short region is an indication of false positives, perhaps caused by small insertions or deletions. GATK's protocols suggest that if we observe two SNPs within 10 bp, the likelihood of a false positive is high. A good common reference for comparison is the 1000 Genomes Project data. The Kolmogorov–Smirnov test can be used to compare the SNP density difference between an observed data set versus the 1000 Genomes Project data.

Cycle bias happens in a heterozygous position when one of two alleles in the supporting reads lies heavily at the beginning or end of the reads. As we already described during the section on raw data quality control, the beginning and ending parts of Illumina's read are more prone to lower quality, thus giving them a higher chance of containing a false-positive SNP. GATK's local realignment procedure can theoretically eliminate some of those false-positive SNPs if insertions or deletions are involved at the beginning or the end of the reads. The average cycle of an SNP is defined as $(\sum_{i=1}^{N} C_i)/N$, where N is the total number of reads that support the alternative allele of this SNP, and C_i denotes the cycle of the SNP on the read. For example, if an SNP is supported by three reads of length 50 bp, and if the SNP happens at the first, first and second cycle on each of these reads, respectively, the average cycle would be $(1+1+2)/3 = 1.3$, which is close to the beginning of the reads, indicating that this SNP may be an artefact. Some aligners such as BWA and Bowtie 2 have soft clip abilities, which can significantly reduce the effect of cycle bias. An alternative method to avoid cycle bias is to perform manual trimming based on read quality before alignment. One artefact of exome capturing is strand imbalance, which is sometimes considered a quality concern. Strand imbalance is the distribution of forward and reverse strands, which can be heavily uneven at many positions, especially those close to the boundaries of target regions. This phenomenon exists for positions both inside and outside target regions, although it is more extreme outside the target regions. In extreme situations, all reads can be on the same strand. It has been shown that strand imbalance has no effect on SNP calling quality [5], so we do not suggest that it be used as a quality control check.

Somatic mutation is harder to detect than SNPs, owing to the involvement of two samples rather than just one. The quality control for somatic mutation is also considerably harder than for SNPs. The common method of identifying somatic mutations in cancer is to compare sequences between paired normal control and tumour samples. If we observe alternative alleles at a genomic position in the tumour but not in the matched normal control at the same position, we assume that this is an acquired somatic mutation in the tumour. If we observe an alternative allele at a genomic position in the normal control but not in the tumour, we call it loss of heterozygosity, again assuming that the somatic variation has formed in the tumour sample, not the normal control. One common strategy used [37–39] to identify somatic mutations is to use the SNP caller to first determine the genotypes of the tumour and control pairs and to then compare the two. Such an approach has certain limitations. First of all, the threshold of mutation for a somatic mutation may be significantly different from that of an SNP. For example, for SNPs, approximately 50% of the reads should support the alternative allele. However, for a somatic mutation, depending on the type of normal control samples used, the expected percentage of mutated reads might differ significantly from 50%. If a blood sample is used as a control, we expect to observe germ line mutations only, whereas if normal tissue adjacent to the tumour is used as a control, the reads observed might represent a mixture of tumour and normal tissues owing to tumour contamination, which can cause the SNP callers to make a false heterozygous inference. Conversely, it is possible that the tumour sample is contaminated by the normal tissue. Programs such as MuTect [34] can adjust for tumour percentage within the sample. However, even after purification procedures such as microdissection, the tumour percentage cannot be estimated accurately. Lower tumour concentration in the sample might cause the sequencer to sequence an insufficient number of reads to support a heterozygous call by current SNP callers. Thus detection of somatic mutations from SNP callers is not recommended. Variant callers based on empirical allele count are more suitable for somatic mutation detection. By bypassing the step of inferring a genotype, we can effectively detect a small percentage of mutations that might not be detectable by an SNP caller and filter out potentially wrong heterozygous inferences from normal controls. The majority of the quality control measures we have described for SNPs also work for somatic mutations. For example, because it is highly

unlikely to observe two non-synonymous mutations in one coding gene, multiple non-synonymous somatic mutations within a single gene raise quality concerns.

There are currently few dedicated analysis tools focusing on variant calling quality control. QC3 (Table S3) is a newly developed quality control tool that targets all three stages of NGS data processing. For a variant call format, QC3 computes the Ti/Tv ratio, the non-reference homozygosity ratio and checks for the possibility of cross-sample contamination.

CONCLUSION

In this review, we have discussed quality control procedures and parameters in extensive detail at each of the three stages of NGS data processing: raw data, alignment and variant calling. We emphasized the importance of conducting quality control checks at all three stages rather than just partially at one or two stages. The overall work flow of our suggested quality control procedures can be seen in Figure 4. The quality control procedures and parameters outlined in Figure 4 can serve as a road map for analysing NGS data thoroughly.

Another quality control method closely related to those discussed in this review is the quality control of library preparation. There are multiple quality control steps during a standard exome sample preparation. The pre-capture DNA libraries undergo a confirmation of size by running an aliquot of the library on the Agilent Bioanalyzer 2100 High Sensitivity Chip, Agilent TapeStation or a 2% gel. The concentration of a pre-capture library is determined using fluorometry such as Picogreen or Qubit for higher accuracy than using a spectrophotometer. After the exome capture hybridizations, the quality control steps are repeated with the addition of using quantitative polymerase chain reaction for accurate molarity. Accurate molarity of each library is essential for balanced read counts in a pool as well as the proper target cluster densities.

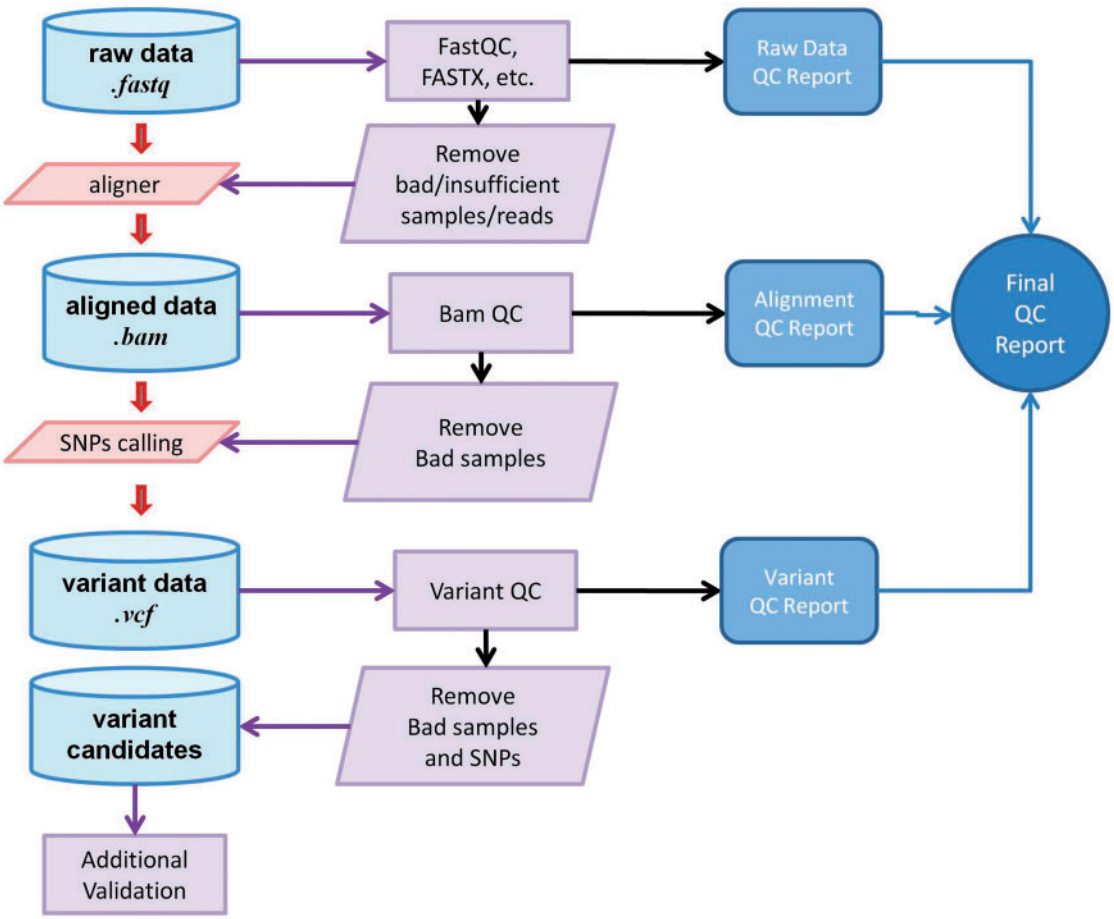

Figure 4: Overall workflow of quality control in DNA sequencing data.

The above mentioned exome library quality control steps are limited because they only address the physical features of the library, the size range and the quantity, but do not measure important features of the library composition. These can only be measured with sequencing and often the true quality of an exome library is not known until the sequencing of the library is complete. With the advent of smaller benchtop sequencers like the Illumina MiSeq or Ion Torrent PGM, preliminary sequencing can be done to gain further information on the quality of a sequencing library or a sequencing library pool. A recent publication also addressed this issue and introduced an empirical Bayesian method to estimate the diversity of a sample and to give information on the amount of data needed for a target coverage level [40]. One of the supported applications of the Illumina MiSeq is sequencing library QC, and features such as the cluster density, library complexity, percent duplication, GC bias and index representation are determined before sequencing the samples at greater depth on an Illumina HiSeq.

Quality control at one stage can have some effects on the choice of the data processing parameters at the next stage. For example, if the quality control on the raw data shows a more-than-usual base quality deterioration, choosing a higher-quality threshold for read trimming (aln–q) in BWA may be recommended. Another example would be if the capture efficiency is low, then limiting the variant calling region to only the capture regions can significantly shorten the calling time. However, in general the results of quality control on one stage will not dramatically affect the best choice of the processing parameters of next stage.

There are many error correction tools for raw data such as Musket [41], HiTEC [42] and SHREC [43]. These tools aim to correct sequencing errors in the raw data. However, they are most useful for *de novo* assembly rather than for alignment against a known reference, the topic of this review. The goals of most re-sequencing studies are to identify variants (SNPs), and performing correction at the raw data level will remove many true SNPs because it is not practical to distinguish between true SNPs and the sequencing errors that these tools are designed to correct. On the other hand, the *de novo* assembly process will benefit by removing both the SNPs and sequencing errors because the goal of *de novo* assembly is to create a consensus sequence.

There are many different alignment tools, and each aligner has its own set of parameters. There should not be a significant influence of the alignment parameters (within reasonable bounds) on how an investigator conducts quality control based on the different parameters chosen. There may be a small difference in the results obtained. For example, allowing more mismatches in the alignment will increase the number of aligned reads in both the capture and non-capture regions. On the other hand, if fairly unreasonable parameters are chosen, such as no mismatch allowed, then there will be no SNPs identified and fewer reads aligned.

Quality control can be a double-edged sword. A good balance between sensitivity and specificity is hard to reach. Depending on the goal of the study, the threshold of the quality control parameters needs to be adjusted accordingly.

Furthermore, there are certain false-positive results that can still evade our quality control efforts even if we do perform the most thorough quality control protocol. Thus, for high-impact studies, the use of additional methods such as Real time, polymerase chain reaction, Sequenom and Sanger sequencing to validate the most important findings independently from the high-throughput sequencing methods is highly recommended.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Quality control is an important component of NGS sequencing analysis.
- There is heavy focus on quality control on raw NGS data.
- Performing quality control only on raw data can result incorrect conclusion.
- Quality control should be performed for NGS DNA sequencing data at three different stages: Raw data, alignment and variant calling.

References

1. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
2. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012;**7**:e30619.
3. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;**27**:863–4.
4. Zhou Q, Su X, Wang A, *et al*. QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One* 2013;**8**:e60234.

5. Guo Y, Long J, He J, *et al*. Exome sequencing generates high quality data in non-target regions. *BMC Genomics* 2012;**13**:194.
6. Ewing B, Hillier L, Wendl MC, *et al*. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998;**8**:175–85.
7. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998;**8**:186–94.
8. Liu Q, Guo Y, Li J, *et al*. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* 2012;**13(Suppl. 8)**:S8.
9. Teer JK, Bonnycastle LL, Chines PS, *et al*. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 2010;**20**:1420–31.
10. Yi X, Liang Y, Huerta-Sanchez E, *et al*. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 2010;**329**:75–8.
11. Ng SB, Buckingham KJ, Lee C, *et al*. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; **42**:30–5.
12. Samuels DC, Han L, Li J, *et al*. Finding the lost treasures in exome sequencing data. *Trends Genet* 2013. http://www.cell.com/trends/genetics/abstract/S0168-9525%2813%2900127-3 (19 September 2013, date last accessed).
13. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.
14. DePristo MA, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8.
15. Durbin RM, Altshuler DL, Abecasis GR, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
16. Guo Y, Li J, Li CI, *et al*. The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 2012;**13**:666.
17. Bainbridge MN, Wang M, Wu Y, *et al*. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol* 2011;**12**:R68.
18. Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 1998;**46**:409–18.
19. Guo Y, Cai Q, Samuels DC, *et al*. The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. *Mutat Res* 2012; **744**:154–60.
20. Lanave C, Tommasi S, Preparata G, *et al*. Transition and transversion rate in the evolution of animal mitochondrial DNA. *Biosystems* 1986;**19**:273–83.
21. Lynch M, Sung W, Morris K, *et al*. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 2008;**105**:9272–7.
22. Morton BR, Bi IV, McMullen MD, *et al*. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* 2006; **172**:569–77.
23. Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**:745–55.
24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**: 1754–60.
26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
27. Li H. Improving SNP discovery by base alignment quality. *Bioinformatics* 2011;**27**:1157–8.
28. Lee H, Graham JM,Jr, Rimoin DL, *et al*. Exome sequencing identifies PDE4D mutations in acrodysostosis. *Am J Hum Genet* 2012;**90**:746–51.
29. Michot C, Le Goff C, Goldenberg A, *et al*. Exome sequencing identifies PDE4D mutations as another cause of acrodysostosis. *Am J Hum Genet* 2012;**90**:740–5.
30. Salzer E, Daschkey S, Choo S, *et al*. Combined immunodeficiency with life-threatening EBV-associatedlymphoproliferative disorder in patients lacking functional CD27. *Haematologica* 2013;**98**:473–8.
31. Koboldt DC, Zhang Q, Larson DE, *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**:568–76.
32. Larson DE, Harris CC, Chen K, *et al*. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 2012;**28**:311–7.
33. Guo Y, Li J, Li CI, *et al*. MitoSeek: extracting mitochondria information and performing high throughput mitochondria sequencing analysis. *Bioinformatics* 2013;**29**:1210–1.
34. Cibulskis K, Lawrence MS, Carter SL, *et al*. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9.
35. Degner JF, Marioni JC, Pai AA, *et al*. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 2009;**25**:3207–12.
36. Guo Y, Samuels DC, Li J, *et al*. Evaluation of allele frequency estimation using pooled sequencing data simulation. *ScientificWorldJournal* 2013;**2013**:895496.
37. Yan XJ, Xu J, Gu ZH, *et al*. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat Genet* 2011; **43**:309–315.
38. Nikolaev SI, Rimoldi D, Iseli C, *et al*. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet* 2012;**44**:133–9.
39. Vissers LE, Fano V, Martinelli D, *et al*. Whole-exome sequencing detects somatic mutations of IDH1 in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). *Am J Med Genet A* 2011;**155A**:2609–16.
40. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. *Nat Methods* 2013;**10**:325–7.
41. Liu Y, Schroder J, Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* 2013;**29**:308–15.
42. Ilie L, Fazayeli F, Ilie S. HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics* 2011;**27**: 295–302.
43. Schroder J, Schroder H, Puglisi SJ, *et al*. SHREC: a short-read error correction method. *Bioinformatics* 2009;**25**:2157–63.