APPLICATIONS OF NEXT-GENERATION SEQUENCING — OPINION

The role of replicates for error mitigation in next-generation sequencing

Kimberly Robasky, Nathan E. Lewis and George M. Church

Abstract | Advances in next-generation sequencing (NGS) technologies have rapidly improved sequencing fidelity and substantially decreased sequencing error rates. However, given that there are billions of nucleotides in a human genome, even low experimental error rates yield many errors in variant calls. Erroneous variants can mimic true somatic and rare variants, thus requiring costly confirmatory experiments to minimize the number of false positives. Here, we discuss sources of experimental errors in NGS and how replicates can be used to abate such errors.

The emergence of next-generation sequencing (NGS) has revolutionized the study of genetics and provided valuable resources for other scientific disciplines. As NGS becomes more widely accessible, its use has extended beyond basic research and into broader clinical contexts. It is therefore increasingly important to account for the errors that arise in the sequencing process. These errors can stem from the bioinformatic analysis[1] and from experimental steps[2,3] (which can often be mitigated through the use of replicate experiments).

The use of replicates permeates almost all scientific disciplines. However, in NGS, many researchers use sequencing read depth and bioinformatic filters to address errors in lieu of biological replication. This practice is understandable, given that replicates can substantially increase study costs. However, sequencing costs have decreased markedly[4], and now is the time to re-evaluate the value of replication in sequencing studies.

In this Perspective article, we discuss sources of errors in sequencing and the nascent use of replication in published high-throughput sequencing efforts. In addition, we show how biological replicates can be used to reduce sequencing errors. In particular, we demonstrate that replicates can be used to assess the specificity and the sensitivity of sequence variant-calling methods in a manner that is independent of the algorithms and the chemistry that are used to call variants, thereby guiding the appropriate selection of quality score thresholds.

Experimental errors in NGS

Technological advances and the digital nature of DNA are helping to achieve highly accurate genome sequences. However, sequencing methods are imperfect. NGS applications — such as whole-genome sequencing, targeted capture, high-throughput RNA sequencing (RNA-seq) and chromatin immunoprecipitation followed by sequencing (ChIP–seq) — are prone to errors that result in miscalled bases, thus causing misalignment of short reads and mistakes in genome assembly. Reported claims of sequencing base call accuracy for leading NGS technologies greatly vary, which range from one error in one thousand nucleotides (99.9%)[5] to one error in ten million nucleotides (99.9999%)[6]. Even for methods that have the lowest reported error rates, the absolute numbers of miscalled genomic variants remain unwieldy — there might be thousands of false-positive variants in a fully sequenced human genome. Furthermore, false-positive errors are mistaken as rare and somatic variants, thereby obfuscating true variants of clinical interest. Known sources for experimental errors can be grouped by their occurrence in the sequencing workflow; that is, during sample preparation, library preparation, or sequencing and imaging (FIG. 1a; BOX 1).

Sample preparation. Sequencing errors and biases can arise from sample degradation and contamination during sample isolation and preservation. For example, during sample preservation, formalin fixation causes degradation and nucleotide changes[7,8]. Moreover, inadequate amounts of high-quality genomic material can increase amplification errors and decrease sequencing read depth[9]. Finally, contamination poses a challenge when non-tumour cells mask oncogenic somatic variants[10] or when exogenous DNA interferes with calls of homozygosity or heterozygosity[11].

Library preparation. Errors also arise during sequencing library preparation, which leads to uneven coverage, sequence changes and interruption of sequence tags. DNA fragmentation can produce length biases, which subsequently causes preferential amplification[12]. Library amplification is subject to unmeasured primer biases, such as primer bias in multiple displacement amplification (MDA)[13], mispriming in PCR target enrichment[14] and incorporation of sequence errors during both clonal amplification and PCR cycling[15]. When barcodes, adaptors and other pre-defined sequence tags are added to the fragments being sequenced, disruption and inadequate tag design can result in cross-contamination of data sets, read loss and decreased read quality[2,16]. Chimeric reads can also arise in long-insert paired-end libraries[17] and potentially confound variant calls and assembly efforts.

Sequencing and imaging. Current NGS platforms[3] have sequencing and imaging error types that are specific to the platforms[18]. For example, substitution errors can arise in platforms such as Illumina and SOLiD when incorrect bases are introduced during clonal amplification of templates. Furthermore, Illumina has shown a sequence-specific error profile[19] that possibly arises from either single-strand DNA folding or sequence-specific alterations in enzyme preference. The single-molecule, real-time (SMRT) platform of Pacific Bioscience yields long single-molecule reads that are subject to false insertions and deletions (indels) from non-fluorescing nucleotides[20,21]. Pyrosequencing (for example, Roche 454 platforms) and semiconductor sequencing (for example, Ion Torrent) have difficulty in counting homopolymer stretches, which results in carry-forward insertion and deletion errors[22].

Experimental errors pose challenges in applications for which accuracy is crucial, such as in detection of somatic mosaicism[23,24] and in other clinical applications. Errors are often addressed by increasing sequencing read depth but can also be mitigated by careful barcoding strategies[25], replicates,

orthogonal sequencing technologies[26] and prior knowledge of variants[27]. Together, these approaches can help to overcome variations in experimental conditions, stochastic fluctuations and systematic biases.

Replicates and experimental errors

Many applications of NGS — for example, the detection of rare causal variants and somatic variants, and clinical applications — require high fidelity in sequencing, which necessitates confirmatory experiments, such as Sanger sequencing. The standard validation methods that are used for confirmation tend to be costly and labour intensive, and lower-cost alternatives are therefore needed.

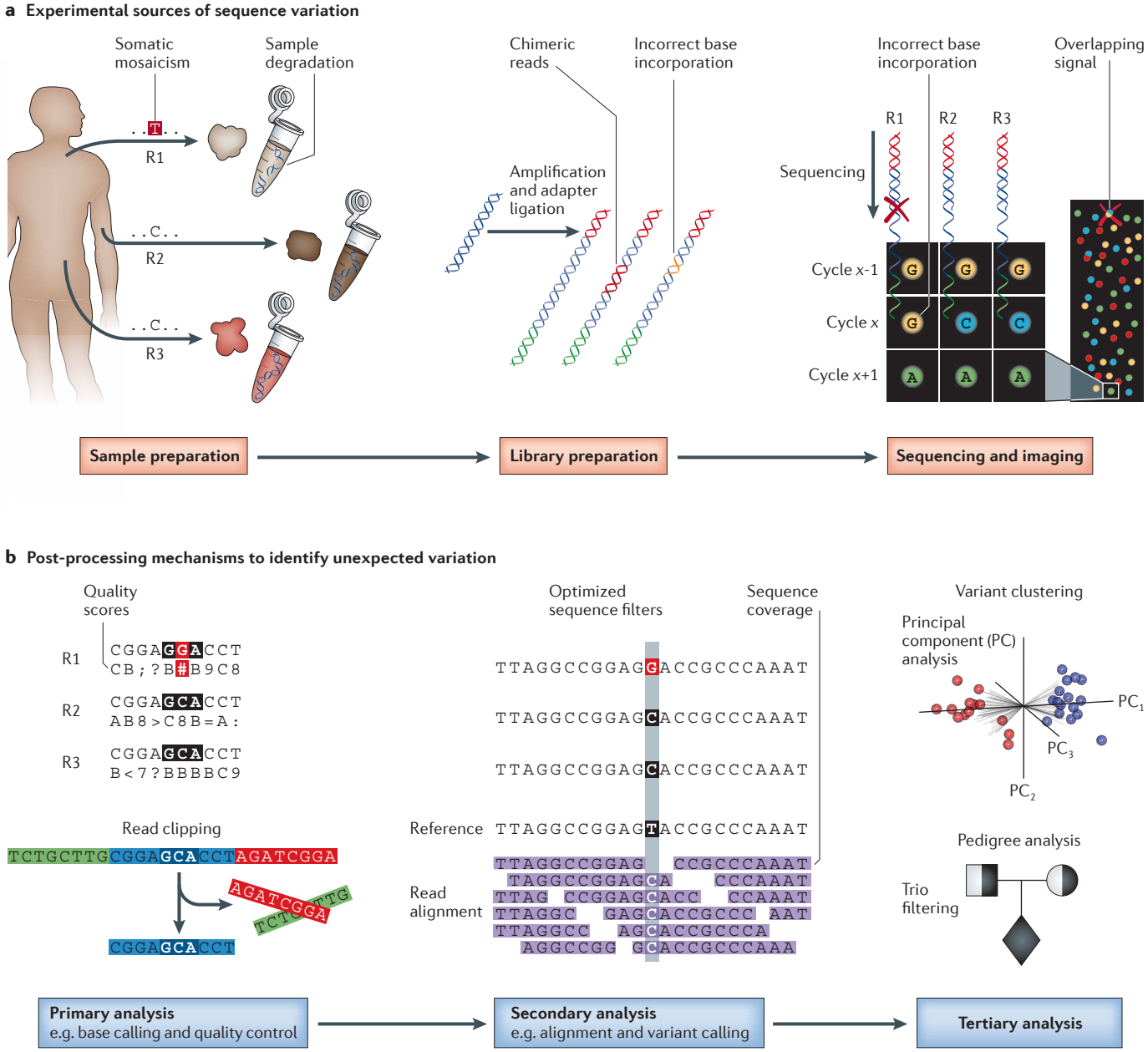

Figure 1 | **Sources of and tools to cope with unexpected or erroneous variants.** Sequencing experiments involve many steps from sample acquisition to final data analysis, and a major challenge in the process stems from the emergence of unexpected or erroneous variants. Sequencing pipeline and sources of errors are shown; R represents a replicate. **a** | These variants can include legitimate somatic mosaicism and rare oncogenic variants. Additionally, many erroneous sequence variants arise during experimental steps, for example, through sample degradation, PCR amplification errors and base-calling errors. **b** | Several analytical tools and post-processing mechanisms are often used for separating true variation from false sequence variants. These include indicators of data quality (for example, base call and mapping quality scores) and the choice of filters that is informed by these indicators. Additional tertiary analyses can also highlight systematic biases through clustering methods and possible false-positive variants by accounting for Mendelian inheritance patterns[57]. Throughout the sequencing and post-processing pipeline, the use of replicated sequencing experiments can help to mitigate the effect of erroneous variants from the experimental steps and to inform the choice of post-processing filters. Thus, greater accuracy of germline variant detection can be attained, and improved sensitivity can be achieved for true somatic variation.

An approach that holds promise uses the tried-and-true scientific method of replication to mitigate user errors, stochastic differences and other sources of experimental errors. Different types of replication are described below, including sequencing read depth, and technical, biological and cross-platform replication.

Sequencing read depth. The most straightforward approach to improve sensitivity and accuracy in sequence variant calls is to increase sequencing read depth[28,29]. By increasing the number of short reads, one can improve variant calling on easily sequenced regions. Consequently, one can reduce the number of missed true variants (that is, false negatives) and sometimes the number of true non-variants that are incorrectly detected as variants (that is, false positives). However, merely increasing sequencing read depth cannot ameliorate issues that arise from the widespread batch effect phenomenon[30] and many other error types that are introduced in the experimental process. Thus, increased sequencing read depth is not necessarily an adequate proxy for biological replication and is limited in its ability to mitigate errors.

Technical replicates. The frequency of certain error types can be reduced through technical replication. We define technical replication as the repeat analysis of the exact same sample. For example, technical replicates were used with monozygotic twins, and the data showed higher intra-individual correlations than inter-individual correlations[31]. In another example[6], many technical replicate pools were sequenced and each contained dilute DNA. Pools containing haplotypes with incongruent base calls that were suspected as amplification errors were discarded, and the sequence quality was significantly improved.

Biological replicates. We define biological replication as the preparation and the analysis of multiple biological samples under the same conditions from the same host. Biological replicates in genome sequencing can be used to assess the efficacy of various bioinformatic filters[32]. Additional benefits over technical replicates include the identification of rare somatic mosaicism and of differences in transcript abundance. Somatic mosaicism can arise from mutations that occur from mutagens and other causes[24]. Biological replicates can indirectly help to uncover somatic mutations in complex and heterogeneous tumours when they are used to achieve the 'normal' baseline sequence in tumour–normal pairs.

Cross-platform replicates. Each sequencing platform introduces unique biases and error types. Thus, integrating sequencing data from different technologies can further mitigate errors. For example, sequencing DNA samples that were taken from both the blood and saliva on two different platforms — Illumina and Complete Genomics — resulted in 88.1% concordance of single-nucleotide variants (SNVs) across replicates[33]. Validation rates for variants that were called on both platforms were higher than variants that were not. In another study, sequencing on three platforms — Illumina, Roche 454 and SOLiD — showed 64.7% concordance[5]. This disparity could result from multiple experimental error sources and from differences in downstream bioinformatic processing. Cross-platform replicates greatly reduce the number of false-positive variants, but the different biases from each sequencing platform may cause many true variants to be overlooked when cross-platform replicates are compared.

Reducing errors and replicates

As sequencing further permeates science and medicine, replicates will be invaluable to researchers and clinicians alike. Current efforts in sequencing error mitigation mainly rely on filtering strategies, including filtering for sequencing read depth, base call quality, short-read alignment quality, variant call quality, known variants, strand bias, allelic imbalance and sequence context[10,21,25,27,34–36]. All of these post-processing techniques help to reduce uncertainty in the final genotyping variant call (FIG. 1b).

Box 1 | **Experimental sources of errors in sequencing**

The importance and the relative effect of each error source on downstream applications depend on many factors, such as sample acquisition, reagents, tissue type, protocol, instrumentation, experimental conditions, analytical application and the ultimate goal of the study. Sequencing errors can stem from any time point throughout the experimental workflow, including initial sequence preparation, library preparation and sequencing. Some examples are listed below.

Sample preparation

- User errors; for example, mislabelling
- Degradation of DNA and/or RNA from preservation methods; for example, tissue autolysis, nucleic acid degradation and crosslinking during the preparation of formalin-fixed, paraffin-embedded (FFPE) tissues[8,87,88]
- Alien sequence contamination; for example, those of mycoplasma and xenograft hosts[89]
- Low DNA input[9]

Library preparation

- User errors; for example, carry-over of DNA from one sample to the next and contamination from previous reactions[90]
- PCR amplification errors[9]
- Primer biases; for example, binding bias, methylation bias, biases that result from mispriming, nonspecific binding and the formation of primer dimers, hairpins and interfering pairs, and biases that are introduced by having a melting temperature that is too high or too low[91,92]
- 3′-end capture bias that is introduced during poly(A) enrichment in high-throughput RNA sequencing[93]
- Private mutations; for example, those introduced by repeat regions and mispriming over private variation[94]
- Machine failure; for example, incorrect PCR cycling temperatures[15]
- Chimeric reads[2,17]
- Barcode and/or adaptor errors; for example, adaptor contamination, lack of barcode diversity and incompatible barcodes[16,95]

Sequencing and imaging

- User errors; for example, cluster crosstalk caused by overloading the flow cell[96]
- Dephasing; for example, incomplete extension and addition of multiple nucleotides instead of a single nucleotide[3]
- 'Dead' fluorophores, damaged nucleotides and overlapping signals[20]
- Sequence context; for example, GC richness, homologous and low-complexity regions, and homopolymers[19,97,98]
- Machine failure; for example, failure of laser, hard drive, software and fluidics
- Strand biases[97]

Bioinformatic filtering techniques can be optimized using technical, biological and cross-platform replicates to improve their specificity and sensitivity[32]. For example, optimal quality score thresholds for each filter may be selected using replicate genome sequences. An individual human genotype has ~3 million variants[36]; however, variant callers can predict >20 million variants of differing quality per genome, which mainly result from mismapped short reads[37], mosaicism and sequencing errors. Consequently, thresholds are chosen to limit the variants called in the individual's genotype. Ideally, these thresholds are chosen with experimental confirmation[38], but this can be costly. We assert that replicates can abet bioinformatic filtering and reduce the number of variants that require validation, thereby improving the quality of the sequence that is being mapped or assembled.

To illustrate this, we use biological replicates to carry out a simple analysis for assessing the reliability of single-nucleotide substitution calls (FIG. 2). For genotyping, the number of replicates should be chosen to attain adequate statistical power at the loci in question. However, in this case, we seek a set of probable false positives that stem from experimental errors, which requires only three replicates for a voting majority. For the replicates, we obtained sequence data from three distinct tissue samples of participant PGP1 in the Personal Genome Project[39] (see Supplementary information S1 (box)).

Loci in which one or more replicates contained a SNV were identified. Briefly, SNV loci are known as concordant when all replicate variant calls agree[40] and discordant when other replicates differ from the target replicate. Thus, concordant loci represent true-positive variants, and discordant loci signal false-positive variants. See Supplementary information S1 (box) for precise definitions of concordance and discordance, for details on choosing a target replicate and for implementation details.

Once discordant variants (potential false positives) and concordant variants (potential true positives) have been separated from each other, metrics of variant call confidence (for example, quality scores and read depth) are used to rank-order the target variants. Using the ranked sets, one can plot the accumulation rate for both concordant and discordant variants with decreasing score stringency in a representation that is similar to a receiver–operator characteristic (ROC) curve (see Supplementary information S2 (box) for methods and source code). Thus, thresholds for variant call quality scores can be chosen to maximize the proportion of all concordant variants that are seen either at or below a particular threshold relative to the proportion of all discordant variants. This analysis (FIG. 3) suggests that, although adequate sequencing read depth across the genome is essential[28,29], it is not the best measure of reliability of a specific variant call at a particular locus. Indeed, sequencing read depth at a particular locus is an inferior filter when it is compared with error-model-based quality scores. We found that this holds true for quality scores that are computed by software packages which process genomic[35] and expression[27] data. Even after removing regions that have abnormally high read depths (that is, regions that are enriched for misalignment errors in low-complexity sequences[37]), the quality scores that are considered here still outperform read depth as a filter for sequencing errors.

In addition to comparing disparate error-model-based quality scores, this approach can be used to evaluate the effect of varying quality score thresholds for a specific data set of interest. For example, sensitivity of a particular threshold can be evaluated by considering the false-negative rate, as estimated by the number of concordant variants that are lost as a result of applying the threshold.

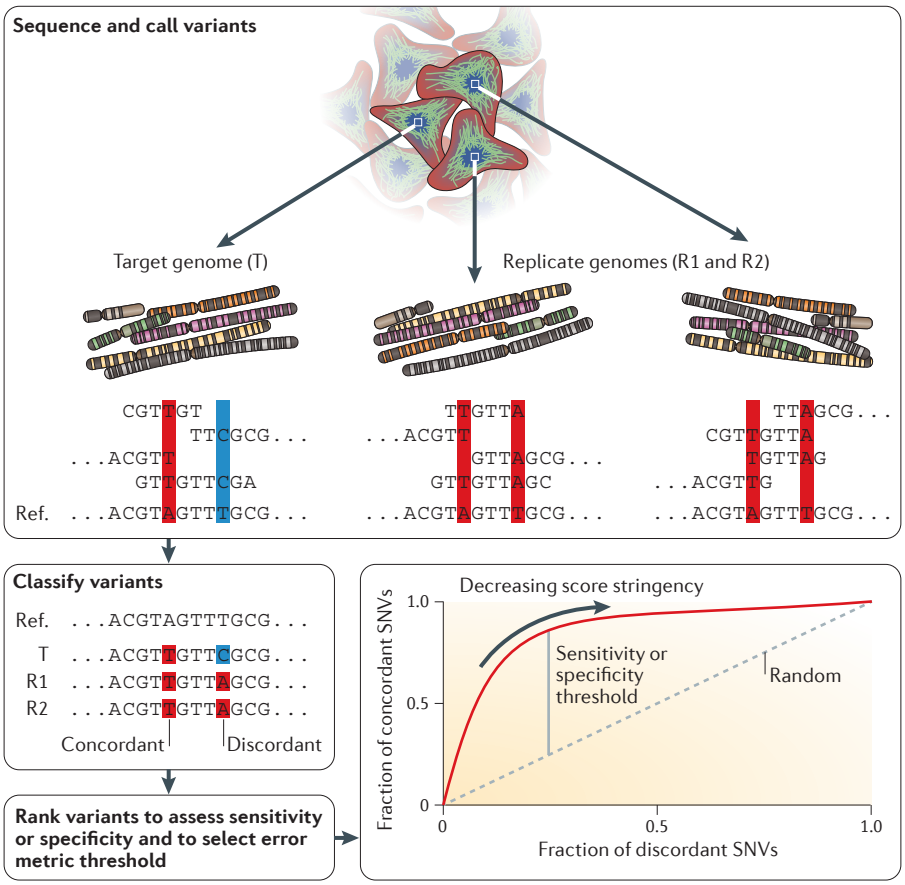

Figure 2 | **Platform-independent method for choosing quality score thresholds.** Single-nucleotide variants (SNVs) are called for all replicates and then classified either as concordant if the variant calls agree among the replicates or as discordant if they differ. Variants are then ranked in order by the desired metric (for example, quality scores) and plotted in a graph that is similar to a receiver–operator characteristic curve; that is, the cumulative distributions of concordant and discordant variants are plotted from left to right as the stringency of the confidence score of interest decreases. Ref, reference sequence.

Post-processing errors in NGS

Even with the use of replicates, some types of errors cannot be addressed without further technological advances and improvements in bioinformatic processing. For example, indels[41], paralogues and other repetitive sequences[42] often confound NGS short-read alignment[43,44], which results in mismapped reads and, ultimately, variant call errors. Other sources of errors can arise from limitations in software and configuration during secondary analysis, including read clipping and filtering[45], allelic bias measurement[46] and variant call confidence

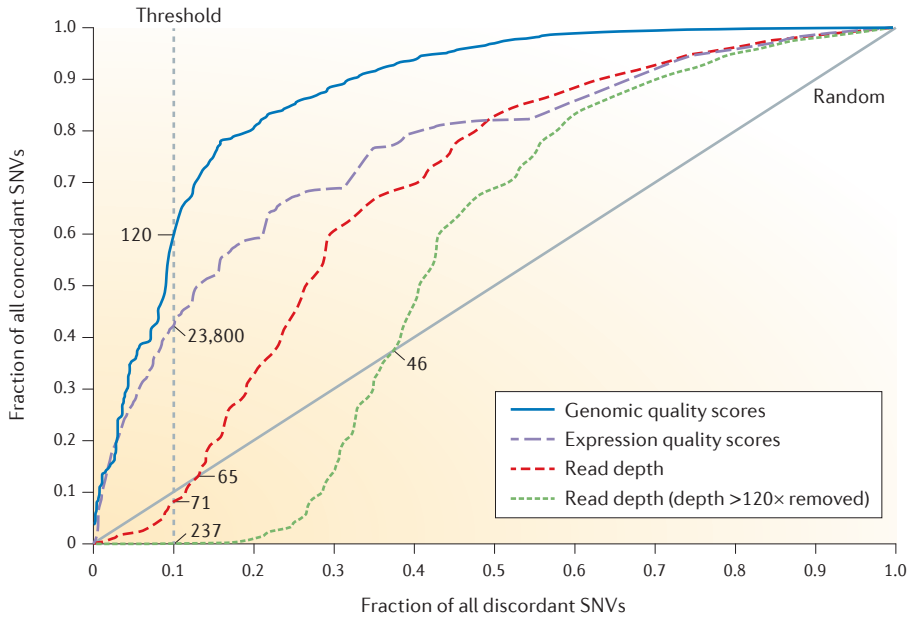

Figure 3 | **An example application of plotting replicate scores to assess filter efficiency.** The efficiency of different variant call filter metrics can be evaluated by plotting replicate-based single-nucleotide variant (SNV) concordance and discordance in a manner that is similar to a receiver–operator characteristic curve. As one goes from left to right on the plot, the quality score that has been ranked in order is reduced in stringency, and the fractions of retained concordant and discordant variants increase. Thus, this curve quantifies the proportion of reliable data (that is, concordant SNVs) that are retained and the proportion of low-confidence data (that is, discordant SNVs) that are discarded as a consequence of variable quality score cutoffs. For the genomes used in our analysis, this graph indicates that filtering variants solely on the basis of locus read depth is inferior to filtering by genomic[35] and expression[27] quality scores[35]. Furthermore, filtering by expression data quality scores is also inferior to filtering by genomic quality scores (which are obtained from Complete Genomics); nevertheless, both of these filters are better than filtering loci by read depth. The read depth curve that excludes outliers (that is, read depth that is higher than the 99.5th-percentile) outperforms the all-inclusive read depth curve. As an example of how to understand the value of a threshold, note that choosing a threshold score of 120 as a measure for the highest quality for the genomic data will include the same fraction of total predicted errors as choosing a threshold quality score of 23,800 for the expression data. Meanwhile, when a similar threshold is chosen for read depth, the efficiency at retaining true variants is worse than that at random. See Supplementary information S2 (box) for a full description of the method.

calculation[47]. These cannot be addressed with replicates alone.

Erroneous variant calls also arise from incomplete reference data. This error type arises when reads are mapped to unfinished reference genomes and transcriptomes, and to drafts that contain misassembled regions[48]. These errors will steadily decrease in frequency as reference genome assemblies and annotations such as GRChr37 (REF. 49) and RefSeq[50] are completed and corrected with each new build release.

Finally, advances in haplotype phasing hold promise not only for reducing amplification errors[6] but also for reducing the causal variation search space. For example, only through accurate haplotype phasing can we begin to discern the difference between two dysfunctional gene copies (that is, a double mutant) and a single normal copy[51]. This difference can have important implications with regard to phenotype and to clinical applications of sequencing. Unfortunately, current mainstream NGS methods do not consistently discern between these two cases. Thus, ad hoc experimental[6,52,53] and computational procedures[54,55] are required to distinguish the haplotypes of diploid cells.

Concluding remarks

In the past decades, scientific and technological advances have provided molecular-level resolution for the inner workings of life. NGS technologies are providing insights into genetic disease associations[56–62], differences in human gut microbiota[63], amino acid essentiality in proteins[64], experimental evolution[65–67], biotherapeutic development[68–72], protein–DNA interactions[73], epigenetics[74], cancer genomics[38,75] and clinical diagnosis[76]. Efforts to find biologically and clinically relevant variants are steadily improving, as algorithmic advances more intelligently filter the large amount of sequence data. For example, priority can be assigned to variants by considering either heritability or variant association in populations[60,77], correcting for gene-specific mutation rates[10], accounting for evolutionary conservation[78–80] and providing network context through systems biology approaches[81–83]. Beyond strictly biological applications, sequencing is also becoming an analytical tool for more esoteric questions, such as recording fluctuations in ion concentrations[84] and even potentially detecting dark matter in astrophysics[85]. However, all these sequencing studies rely on the accuracy of the underlying sequencing experiments.

Here, we have identified sources of sequencing errors and presented a method for addressing the stochastic effects. Additional approaches to address other sources of errors, such as experimental bias and software limitations, are also essential. These approaches include identifying erroneous single-nucleotide polymorphisms that show Hardy–Weinberg disequilibrium[11], masking poor-quality bases[86], phasing and imputing variants in regions that are difficult to sequence or in uncalled regions[54], as well as improved methods for calling of structural variants, copy number variations and indels. Together with these computational approaches, the wise use of replicate genome sequencing will have an increasingly important role in reducing the noise in data processing and in downstream analyses.

Kimberly Robasky was previously at the Program in Bioinformatics, Boston University, Massachusetts 02115, USA; the Department of Genetics, Harvard Medical School, and the Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, Massachusetts 02115, USA. Present address: Expression Analysis, a Quintiles Company, Durham, North Carolina 27713, USA.

Nathan E. Lewis was previously at the Department of Genetics, Harvard Medical School, and the Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, Massachusetts 02115, USA; and the Department of Biology, Brigham Young University, Provo, Utah 84602, USA. Present address: Division of Pediatric Pharmacology and Drug Discovery, University of California, San Diego School of Medicine, La Jolla, California 92093, USA.

George M. Church is at the Department of Genetics, Harvard Medical School, and the Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, Massachusetts 02115, USA.

K.R. and N.E.L. contributed equally to this work.

Correspondence to N.E.L.
e-mail: natelewis3@gmail.com

doi:10.1038/nrg3655
Published online 10 December 2013

Glossary

Barcodes
Known DNA sequences that are appended to the ends of DNA fragments before sequencing for the purpose of pooling samples together to reduce cost.

Base call
Identification of the nitrogenous base (A, G, C or T) that is added to the short read during sequencing.

Batch effect
The statistical bias of indeterminate cause observed in samples that are processed together with the same sample preparation, the same library preparation and the same sequencing experiment.

Homopolymer
A sequence of multiple consecutive identical nucleotides.

Insertions and deletions
(Indels). Variants that are created by either the insertion or the deletion of nucleotides with respect to a matching reference.

Misalignment
The alignment of a sequencing read to an incorrect location on a reference genome. This can occur when reads align equally well to multiple genomic locations owing to indels, repeats and low-complexity regions of the genome.

Multiple displacement amplification
(MDA). A technique that is used for amplifying DNA sequences by synthesizing DNA from random hexamer primers.

Read clipping
Removal of adaptor and barcode sequences or of low-quality bases near read ends following sequencing.

Sequencing errors
Errors that are seen in the base call of short reads from next-generation sequencing technology.

Sequencing read depth
The number of reads that contributes to the variant call at a single location; also known as read depth, fold coverage and depth of coverage. It can also refer to the average read depth across the entire targeted sequence area.

Short reads
Short sequences of nucleotide bases and their respective quality scores that are obtained through next-generation sequencing from longer target sequences.

Somatic mosaicism
Genetic diversity among cells of a single organism.

Substitution errors
Errors that occur when one base is substituted for another during sequencing.

Variant call errors
An accumulation of misaligned reads or of reads with base call errors over a particular locus, which results in that locus being called a variant when it truly matches the reference, and vice versa.

1. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).
2. Kircher, M., Heyn, P. & Kelso, J. Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics* **12**, 382 (2011).
3. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
4. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, 125 (2011).
5. Ratan, A. *et al.* Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS ONE* **8**, e55089 (2013).
6. Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
7. Williams, C. *et al.* A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am. J. Pathol.* **155**, 1467–1471 (1999).
8. Yost, S. E. *et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40**, e107 (2012).
9. Akbari, M., Hansen, M. D., Halgunset, J., Skorpen, F. & Krokan, H. E. Low copy number DNA template can render polymerase chain reaction error prone in a sequence-dependent manner. *J. Mol. Diagn.* **7**, 36–39 (2005).
10. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
11. Leal, S. M. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy–Weinberg equilibrium. *Genet. Epidemiol.* **29**, 204–214 (2005).
12. Walsh, P. S., Erlich, H. A. & Higuchi, R. Preferential PCR amplification of alleles: mechanisms and solutions. *PCR Methods Appl.* **1**, 241–250 (1992).
13. Hutchison, C. A. 3rd, Smith, H. O., Pfannkoch, C. & Venter, J. C. Cell-free cloning using phi29 DNA polymerase. *Proc. Natl Acad. Sci. USA* **102**, 17332–17336 (2005).
14. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
15. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
16. Bystrykh, L. V. Generalized DNA barcode design based on Hamming codes. *PLoS ONE* **7**, e36852 (2012).
17. Koboldt, D. C., Ding, L., Mardis, E. R. & Wilson, R. K. Challenges of sequencing human genomes. *Brief Bioinform.* **11**, 484–498 (2010).
18. Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett.* **340**, 284–295 (2012).
19. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011).
20. Fuller, C. W. *et al.* The challenges of sequencing by synthesis. *Nature Biotech.* **27**, 1013–1023 (2009).
21. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* **14**, 405 (2013).
22. Yang, X., Chockalingam, S. P. & Aluru, S. A survey of error-correction methods for next-generation sequencing. *Brief Bioinform.* **14**, 56–66 (2013).
23. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
24. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genet.* **44**, 642–650 (2012).
25. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
26. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina versus Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* **7**, e30087 (2012).
27. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
28. Ajay, S. S., Parker, S. C., Abaan, H. O., Fajardo, K. V. & Margulies, E. H. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505 (2011).
29. Meynert, A. M., Bicknell, L. S., Hurles, M. E., Jackson, A. P. & Taylor, M. S. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics* **14**, 195 (2013).
30. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Rev. Genet.* **11**, 733–739 (2010).
31. Baranzini, S. E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351–1356 (2010).
32. Reumers, J. *et al.* Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotech.* **30**, 61–68 (2012).
33. Lam, H. Y. *et al.* Performance comparison of whole-genome sequencing platforms. *Nature Biotech.* **30**, 78–82 (2012).
34. Jung, H., Bleazard, T., Lee, J. & Hong, D. Systematic investigation of cancer-associated somatic point mutations in SNP databases. *Nature Biotech.* **31**, 787–789 (2013).
35. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
36. Pelak, K. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet.* **6,** e1001111 (2010).
37. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
38. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
39. Ball, M. P. *et al.* A public resource facilitating clinical use of genomes. *Proc. Natl Acad. Sci. USA* **109**, 11920–11927 (2012).
40. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
41. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
42. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
43. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
44. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
45. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337 (2012).
46. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
47. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
48. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nature Genet.* **45**, 406–414 (2013).
49. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).
50. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
51. Rusk, N. One genome, two haplotypes. *Nature Methods* **8**, 107 (2011).
52. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotech.* **29**, 51–57 (2011).
53. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotech.* **29**, 59–63 (2011).
54. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Rev. Genet.* **12**, 703–714 (2011).
55. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
56. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).
57. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

58. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
59. Chapman, S. J. & Hill, A. V. Human genetic susceptibility to infectious disease. *Nature Rev. Genet.* **13**, 175–188 (2012).
60. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nature Rev. Genet.* **12**, 465–474 (2011).
61. Gibson, G. Rare and common variants: twenty arguments. *Nature Rev. Genet.* **13**, 135–145 (2011).
62. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Rev. Genet.* **11**, 843–854 (2010).
63. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
64. Robins, W. P., Faruque, S. M. & Mekalanos, J. J. Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. *Proc. Natl Acad. Sci. USA* **110**, E848–857 (2013).
65. Conrad, T. M., Lewis, N. E. & Palsson, B. O. Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* **7**, 509 (2011).
66. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
67. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nature Rev. Genet.* **14**, 827–839 (2013).
68. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nature Biotech.* **29**, 735–741 (2011).
69. Lewis, N. E. *et al.* Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nature Biotech.* **31**, 759–765 (2013).
70. Brinkrolf, K. *et al.* Chinese hamster genome sequenced from sorted chromosomes. *Nature Biotech.* **31**, 694–695 (2013).
71. Becker, J. *et al.* Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *J. Biotechnol.* **156**, 227–235 (2011).
72. Kildegaard, H. F., Baycin-Hizal, D., Lewis, N. E. & Betenbaugh, M. J. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr. Opin. Biotechnol.* **24**, 1102–1107 (2013).
73. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Rev. Genet.* **13**, 840–852 (2012).
74. Meaburn, E. & Schulz, R. Next generation sequencing in epigenetics: insights and challenges. *Semin. Cell Dev. Biol.* **23**, 192–199 (2012).
75. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
76. Rios, J., Stein, E., Shendure, J., Hobbs, H. H. & Cohen, J. C. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum. Mol. Genet.* **19**, 4313–4318 (2010).
77. Schneeberger, K. *et al.* SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods* **6**, 550–551 (2009).
78. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Rev. Genet.* **12**, 628–640 (2011).
79. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nature Methods* **10**, 723–729 (2013).
80. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
81. Lewis, N. E. & Abdel-Haleem, A. M. The evolution of genome-scale models of cancer metabolism. *Front. Physiol.* **4**, 237 (2013).
82. Ala-Korpela, M., Kangas, A. J. & Inouye, M. Genome-wide association studies and systems biology: together at last. *Trends Genet.* **27**, 493–498 (2011).
83. Moreau, Y. & Tranchevent, L. C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Rev. Genet.* **13**, 523–536 (2012).
84. Zamft, B. M. *et al.* Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS ONE* **7**, e43876 (2012).
85. Drukier, A. *et al.* New dark matter detectors using DNA for nanometer tracking. *arXiv* 1206.6809 (2012).
86. Hubisz, M. J., Lin, M. F., Kellis, M. & Siepel, A. Error and error mitigation in low-coverage genome assemblies. *PLoS ONE* **6**, e17034 (2011).
87. Macabeo-Ong, M. *et al.* Effect of duration of fixation on quantitative reverse transcription polymerase chain reaction analyses. *Mod. Pathol.* **15**, 979–987 (2002).
88. Kerick, M. *et al.* Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genom.* **4**, 68 (2011).
89. Lin, M. T. *et al.* Quantifying the relative amount of mouse and human DNA in cancer xenografts using species-specific variation in gene length. *Biotechniques* **48**, 211–218 (2010).
90. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. *PCR protocols: a guide to methods and applications* (Academic press, 1990).
91. Wojdacz, T. K., Hansen, L. L. & Dobrovic, A. A new approach to primer design for the control of PCR bias in methylation studies. *BMC Res. Notes* **1**, 54 (2008).
92. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* **96**, 317–323 (2003).
93. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
94. Pont-Kingdon, G. *et al.* Design and analytical validation of clinical DNA sequencing assays. *Arch. Pathol. Lab Med.* **136**, 41–46 (2012).
95. Gogol-Doring, A. & Chen, W. An overview of the analysis of next generation sequencing data. *Methods Mol. Biol.* **802**, 249–257 (2012).
96. Whiteford, N. *et al.* Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics* **25**, 2194–2199 (2009).
97. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotech.* **30**, 434–439 (2012).
98. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).

Acknowledgements

The authors thank T. Gianoulis for her feedback and inspiration, and J. Dupuis, Professor of Biostatistics at Boston University, Massachusetts, USA, for her encouragement and feedback during the nascent stages of replicate analysis. They also thank W. Jones, Global Head of Genomic Bioinformatics, Quintiles, and E. Aronesty, author of the ea-utils FASTQ processing package, for critical review of the manuscript. Some of this work was supported by the US National Institutes of Health grant P50HG005550.

Competing interests statement

The authors declare competing interests: see Web version for details.

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF