Journal of
Cancer Science & Therapy

Blighe et al., J Cancer Sci Ther 2014, 6:10
DOI: 10.4172/1948-5956.1000300

Research Article Open Access

Next Generation Sequencing in the National Health Service England: A Pipeline that Completely Agrees with Sanger

Kevin Blighe, Nick Beauchamp, K Elizabeth Allen, Isabel M Nesbitt, Jennifer Dawe, Darren Grafham and Ann Dalton*

Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield, UK

Abstract

As next generation sequencing (NGS) technology has already become a regular fixture in research, it is now time for clinical environments to also reap the benefits of such technology. Indeed, the rich promise of NGS has the potential to be translated into improved patient care. However, there is still doubt about the widespread use of NGS in clinical diagnostics. Before implementation, there must be consensus on which analytical pipeline to use, with follow-up confirmation of variants with the gold standard: Sanger sequencing.

Here, we present a NGS analytical pipeline that has complete agreement on 341 variants with Sanger sequencing and that is already being used in our clinical diagnostic laboratory in the National Health Service England for the regular screening of inherited, pathogenic variants. Details on our NGS and other services can be found at http://www.sheffieldchildrens.nhs.uk/our-services/sheffield-diagnostic-genetics-service/. Our pipeline broadly follows the 'best practices' guidelines set by the GATK at the Broad Institute, with a novel added approach involving randomly selecting subsets of reads and later merging variants called from each. This allows for false-negatives to be eliminated with a high level of confidence. Moreover, modeling reduced depth of coverage reveals that 30X is the point at which false-positives are eliminated with >99.9% confidence.

Our results allude to a fine balance between read-depth and error, and we believe that our pipeline will increase confidence in NGS and permit its gradual enrollment in clinical diagnostic laboratories.

Keywords: Next generation sequencing; Clinical diagnostics; Sanger agreement; Random read selection

Introduction

When we speak of data in the realm of next generation sequencing (NGS), we typically refer to colossal amounts of data-points that require analysis, filtering, and interpretation [1,2]. Analytical pipelines can be automated such that these prodigious amounts of data are processed with the single push of a button, but this begs the question: What is the most faithful way of analyzing NGS raw data?

Previously, the best way to prove the faithfulness and capabilities of NGS was to find concordance with the gold standard in the clinical diagnostic setting [3-5], i.e., di-deoxy Sanger sequencing [6]. Although it is known that Sanger itself is prone to error [7], it is still regarded as the fundamental method of detecting DNA mutations against which NGS must compare [8]. Indeed, many variant callers and pipelines for NGS data are now in existence [2,9-13], but complete agreement with Sanger remains either elusive or it is something that is no longer sought. Although good agreement with Sanger was recently achieved [3], only 168 variants were observed and the method was not reproducible. Moreover, power analysis has shown that agreement on 300 or more variants is necessary [14], a figure also adopted by the British-based Association for Clinical Genetic Sciences (ACGS, http://www.acgs.uk.com/) in their best practice guidelines. Further, NGS data is still plagued with false-positives and –negatives [15-17] that serve to dampen confidence in its reliability as a mode of diagnosis. As a final issue, there is still very much a lack of consensus on how to analyze NGS raw data, an area in which systematic methods are required [18], with different organizations and commercial ventures applying different filtering and QC thresholds. The community appears undecided on what is or is not a true variant.

Thus, we are set in the tantalizing situation whereby complex, rare, and other diseases are being genetically characterized in the research world [19-25], but as yet no analytical pipeline has been capable of increasing confidence in NGS technology such that it can replace existing clinical diagnostic methods. A critical point in this regard is that the research world can tolerate a certain level of error in results, whereas the demands of a clinical diagnostic service are much higher. It is believed that generating high depth of coverage can boost NGS' capabilities —particularly for somatic mutation detection in cancer [4] — however, it was previously shown that increasing depth actually resulted in more false-positives [26]. Indeed, it seems that there is a fine balance between read-depth and error, and that most of the core issues pertaining to NGS surround the bioinformatic algorithms used to process and filter the raw data [18]. To overcome these issues, many have adopted the strategy of analyzing NGS raw data using multiple variant callers and/or pipelines and then finding a consensus [27], but it was clear that much disagreement still existed between each set of results. Using replicate samples was also recently suggested as a way to minimize error [17], but this is impractical where DNA is limited.

Thus, whilst there is already much confidence in the capability of NGS, barriers remain to its long term use in clinical diagnostics. We therefore set out to develop an analytical pipeline that could eventually be used as the sole method of diagnosis in our laboratory. This required the following:

***Corresponding author:** Ann Dalton, Sheffield Diagnostic Genetics Service, C Floor, Blue Wing, Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield, S10 2TH, UK, Tel: +44 114 271 7014; E-mail: ann.dalton@sch.nhs.uk

Received August 29, 2014; **Accepted** September 27, 2014; **Published** September 29, 2014

Citation: Blighe K, Beauchamp N, Allen KE, Nesbitt IM, Dawe J, et al. (2014) Next Generation Sequencing in the National Health Service England: A Pipeline that Completely Agrees with Sanger. J Cancer Sci Ther 6: 401-405. doi:10.4172/1948-5956.1000300

1) Ensuring complete pick-up of variants, thus eliminating false-negatives
2) Determining thresholds to filter-out false-positives
3) Adhering to data protection legislation and respecting patient privacy
4) Outputting results according to standard nomenclature
5) Broadly, providing all data needed by those interpreting the results in order that informed decisions could be reached

All of these outcomes need to be achieved for a robust clinical analysis pipeline.

Methods

We used the SureSelect (Agilent Technologies, Inc.) wet-laboratory chemistry and protocol for the screening of variants under the umbrella of two main disease areas: connective tissue disorders (CTDs) and glycogen storage diseases (GSDs). We designed custom panels of probes using SureSelect for each of these. Additionally, we used the TruSight™ Cancer Sequencing Panel (Illumina, Inc.) for the screening of hereditary cancers. Both panels differ in their protocols, with TruSight™ having a longer set-up time (~1.5 days compared to 1 day for SureSelect) and utilizing shorter probes (80-mer compared to 120-mer SureSelect probes). All of our NGS was performed with an Illumina MiSeq™ (Illumina, Inc.), whilst our di-deoxy Sanger sequencing was performed using an Applied Biosystems 3730xl DNA Analyzer (Applied Biosystems®). For visualizing NGS data, we used Alamut (Interactive Biosoftware, LLC.) and Integrated Genomics Viewer (IGV) [30,31]. We also used SAP® Crystal Reports® (SAP AG) for tabulating data.

Random read-selection for recovering false-negatives

To perform the validation of our NGS analytical pipeline with Sanger, we used the raw data obtained from 14 patient samples that were sequenced using SureSelect: 7 samples using the CTD custom panel; 7 using the GSD custom panel. We also used the raw data from 19 samples using TruSight™ Cancer Sequencing Panel. We obtained the unfiltered raw data (FASTQ files) for each sample and passed these through the analytical pipeline (Figure 1).

In brief, paired alignment was performed using BWA [32]. We then marked PCR duplicates using Picard (http://picard.sourceforge.net/) and expunged these duplicate reads from the BAM files prior to performing QC. SAMtools [33], BEDTools [34], GATK [2,35], and custom shell commands were used to generate various QC reports, including: alignment and reads on target percentages; coverage at various depths; minimum and maximum read-depth; and a report of all bases falling below a defined threshold (variants called on any of these bases below the threshold are not considered). Any sample that fails QC is discussed in the scientist meeting (Figure 1), with the possibility of repeating the sample in the wet-laboratory.

Post-QC steps involved preparing the data for variant calling. First, candidate indels were identified and then realigned using GATK [2,35]. We then adopted a novel approach, as follows: Random sets of half and quarter reads were extracted from the aligned BAM file using Picard (http://picard.sourceforge.net/), which looks at each read in the BAM file and uses a predefined probability of retaining or discarding the read (in this pipeline, the probabilities were fixed at 0.5 and 0.25 for half and quarter sets of reads). The full, half, and quarter sets were then passed into GATK Haplotype Caller [2,35], with each set of variants

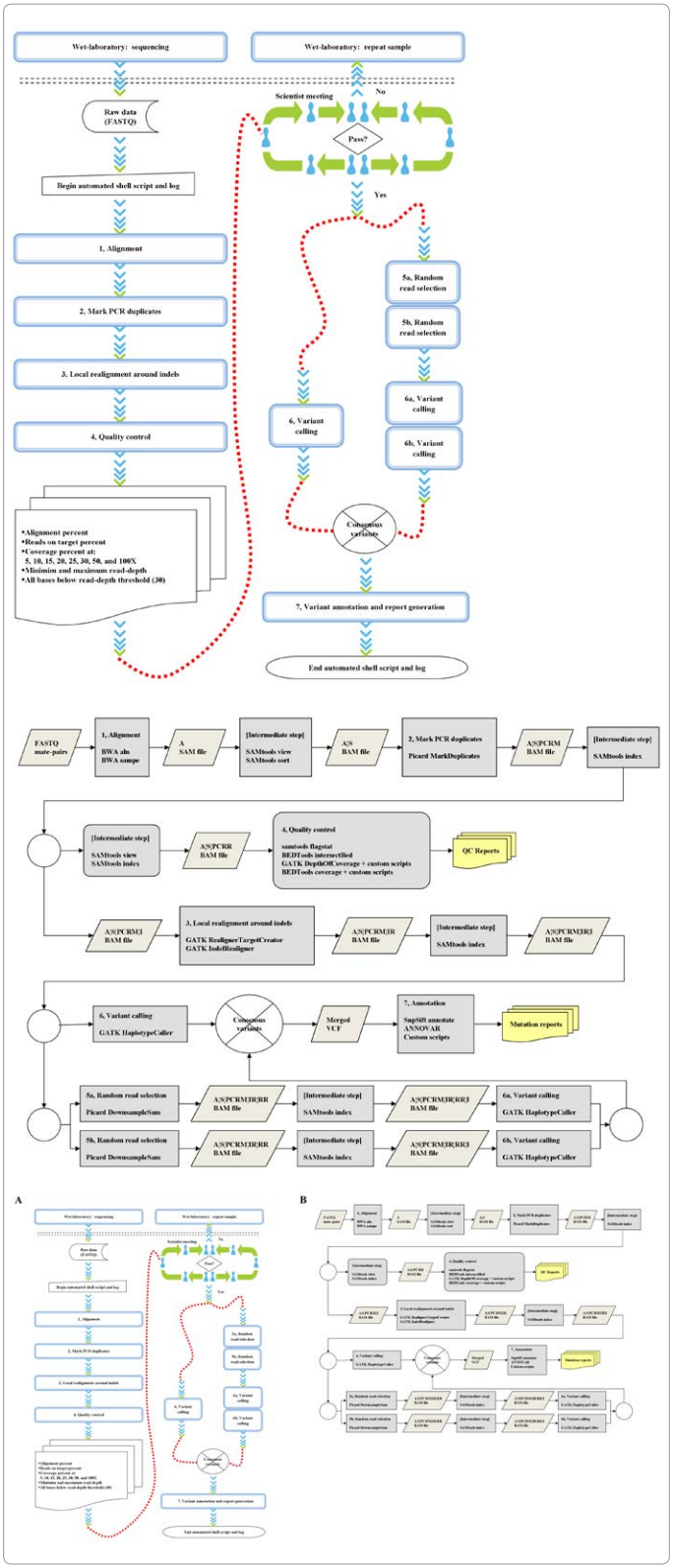

Figure 1: Analytical pipeline diagram. A, broad overview of operations highlighting the interactions between the wet-laboratory and scientists with the analytical pipeline; **B**, all inputs, outputs, and programs used at each step in the analytical pipeline. A, aligned; S, sorted; PCRM, PCR duplicates marked; PCRR, PCR duplicates removed; I, indexed; IR, indels realigned; RR, random read-sampled.

then being merged into a consensus list of variants. The consensus list was annotated using SnpSift [36] and ANNOVAR [37] before being filtered according to each respective BED file regions of interest (ROI) using VCFtools [38]. For the purposes of this work, the pipeline was configured to look up to 150bp into each intronic region for calling and annotating variants. Custom shell commands were again used to present the data in the form of mutation reports. All variants called in the ROI were then manually compared to Sanger sequence results for the same regions. SAMtools [33] was used for intermediate steps throughout the workflow that involved file format conversion, sorting, or indexing.

Modeling reduced depth of coverage to avoid false-positives

For modeling reduced depth of coverage, we used the raw data obtained from three samples that were sequenced using SureSelect: two samples using the CTD custom panel; one using the GSD custom panel (these were samples that had also been previously sequenced by Sanger in our laboratory). We passed the unfiltered raw data (FASTQ files) through our NGS analytical pipeline to completion. We subsequently sampled reads from the aligned BAM files for each and then processed these read libraries through the remainder of the pipeline. This was to simulate reduced depth of coverage. Samplings were performed at levels of 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128, 0.256 and 0.512.

We filtered variants according to our BED file ROI, which matched the primer pools used in each SureSelect kit, configuring the pipeline to again look up to 150bp into each intronic region for the purpose of calling variants. Additionally, for this modeling, as we were searching for an ideal level of coverage to be used as our main QC threshold, we used the existing Phred-scaled quality score as an initial gauge of quality, taking 60 or more as a cut-off (equating to a 1-in-1 million chance of being incorrect). We took the variant caller format (VCF) [38] files that were annotated with dbSNP [39] (build 138) for each sample and its associated read libraries, and used SAP® Crystal Reports® to perform our analyses. When performing the final analyses in modeling, we excluded variants that represented long indels (≥ 20bp) and those that represented the same indel called in the same sample at varying lengths - we also excluded variants called in homopolymeric and repeat regions.

Results

Random read-selection for recovering false-negatives

As high sensitivity with the gold standard is a requirement for any NGS analytical pipeline in a clinical diagnostic laboratory, we sought to compare the level of agreement between our NGS and Sanger sequencing results. In total, we compared results for 341 Sanger-confirmed variants covering 28 genes. These included both exonic and intronic variants, and also single nucleotide variants (SNVs) and insertions-deletions (indels), and were tabulated from the results of 33 patient DNA samples screened in our laboratory for various inherited disorders: glycogen storage diseases, connective tissues disorders, and hereditary cancers. The results are summarized in Table 1. We were able to detect all Sanger-confirmed variants using NGS, i.e., no false-negatives were found. In addition, our NGS analytical pipeline detected 8 intronic variants that were overlooked during Sanger sequence analysis due to poor quality. These were subsequently detected when the Sanger traces were rechecked.

Modeling reduced depth of coverage to avoid false-positives

With confidence that all variants were being detected, we then set-out to find a read-depth 'sweet spot' in order to cope with the uncomfortable amount of false-positives that NGS results can contain. We wanted to identify the lowest read-depth at which a variant could still be detected and at which we could still be confident in the call. To do this, we modeled reduced depth of coverage in samples and found the level of read-depth at which high sensitivity could still be achieved. In total, we analyzed 3,121 variants that were called at any initial level of read-depth across 3 patient samples. Although some of these were common to 2 or 3 samples, we decided to treat each as unique. These variants covered multiple genes across the genome and included variants in both coding and non-coding regions. We sampled the raw reads from these samples at 9 reduced levels, ran each through our analytical pipeline (Figure 1 and Methods), and then compared

Connective tissue disorders		Glycogen storage diseases		Hereditary cancers	
Gene	**Total variants**	**Gene**	**Total variants**	**Gene**	**Total variants**
ALPL	14	*AGL*	19	*APC*	8
COL1A1	21	*GAA*	19	*BRCA1*	22
COL1A2	24	*GBE1*	5	*BRCA2*	32
COL3A1	12	*GYS2*	15	*FANCA*	33
COL5A1	49	*PHKB*	1	*FANCG*	3
COL5A2	14	*PHKG2*	2	*MLH1*	3
CRTAP	5	*PYGL*	11	*MSH2*	4
FKBP10	1				
LEPRE1	5				
PLOD1	1				
PPIB	1				
SERPINF1	11				
SERPINH1	4				
SP7	2				
Total	**164**	**Total**	**72**	**Total**	**105**
Overall total	**341**				

Table 1: Validated Sanger and NGS variants. Validated variants are totaled per gene and disease area under which they were originally detected. Not listed are intronic variants that were missed by Sanger sequencing but detected by NGS.

results for each variant, separately. The lowest read library contained in the region of 4000 reads in each FASTQ mate-pair file, roughly doubling in each library to approximately 1 million reads in the 0.512 library. Indeed, read-depth scaled up with each successive read library (Pearson correlation r^2 0.94).

We then took different cut-off values for read-depth in order to test the sensitivity at each, the question being: assuming that those variants called in the primary 'unsampled' read-set were true variants, which was the lowest read-depth at which we could still detect 100% of these? Using a read-depth of 18 as an initial cut-off, as this was the level of depth of coverage quoted on mutation reports by laboratories within the United Kingdom Genetic Testing Network (UKGTN) (http://ukgtn.nhs.uk/), we found that many variants had a lowest detectable read-depth of much below 18. Indeed, many were still being correctly called from a read-depth of just 2.

There were only 132 variants whose lowest detectable read-depth was greater than this initial cut-off of 18. We individually examined these 132 variants and found that 118 had started with an initial read-depth of below 18 (i.e. in the unsampled read-set); thus, we excluded these on the basis that a read-depth of 18 was the absolute minimum at which we had confidence in a call. This left 3,003 variants out of the original total. For the remaining 14 variants, there was no explanation for their not having been detected below a read-depth of 18. Thus, the sensitivity is 99.5% at a read-depth of 18. We then repeated this analysis at higher cut-offs of read-depth, and attained sensitivities of 99.6%, 99.9%, and 100% at read-depths of 20, 25, and 30, respectively. We therefore decided to use 30 as our threshold for all future analyses using our clinical NGS analytical pipeline.

Discussion

We have developed a NGS analytical pipeline that has complete agreement with Sanger sequencing on 341 variants and that is currently being used in the live diagnostic setting in our laboratory in the National Health Service (NHS) England. Moreover, we have shown how our random read selection approach may give the pipeline greater sensitivity than Sanger in poor quality regions; thus, although complete concordance with Sanger is clearly possible, it may actually be inappropriate when setting clinical analysis standards for NGS in the future.

Our depth of coverage modeling analysis is important. Next generation sequencing data is known to suffer from 'chronic' false-positives; however, as opposed to filtering variants based on Phred-scaled quality scores, genotype likelihoods, etc., we decided to base our filtering solely on read-depth. As a result of our work, we can conclude that we are confident of a variant call made with a position read-depth of 30 or more. Below this, the confidence begins to tail off; however, even at a read-depth of 18, we have a confidence of 99.5%. At lower levels of depth, we noticed a high introduction of false positives, and thus advise others against ever using thresholds below at least 18. However, the issue of false-negatives or 'missed' variants is perhaps the critical finding from this study. Although we employed the widely-used 'best practices' guidelines by the GATK (https://www.broadinstitute.org/gatk/guide/best-practices), with additional steps to include random read selection, we have shown how the use of the same variant caller will result in different variants being called on the same DNA sample, depending on the number of reads present. In addition, we have shown that increasing depth of coverage does not necessarily counteract this issue. Applying the random read selection steps in our analytical pipeline copes with this issue.

Finally, we have shown how our analytical pipeline is suited to the clinical setting for several reasons: Firstly, at no point during the automated analysis is data transmitted outside the domain in which the pipeline is run. This ensures adherence to standards pertaining to data protection and patient privacy (for example, BS7799 in the United Kingdom of Great Britain and Northern Ireland). Secondly, we have conducted comprehensive validation work, including our reduced depth of coverage analysis and our comparison of results with Sanger sequencing. Additionally, we output our variants according to the standards set by the Human Genome Variation Society (HGVS) [40]. Finally, we have shown how the analytical pipeline functions equally on raw data produced from different wet-laboratory chemistries.

To summarize, we have developed a robust NGS analytical pipeline that consistently agrees with the current gold standard in clinical diagnostics: di-deoxy Sanger sequencing. This pipeline is currently automated and is being used in the live diagnostic setting at the Sheffield Diagnostic Genetics Service, part of the Sheffield Children's NHS Foundation Trust (http://www.sheffieldchildrens.nhs.uk/our-services/sheffield-diagnostic-genetics-service/). We encourage other clinical diagnostic laboratories and research groups to test our method.

Acknowledgements

We wish to thank Professor Anne Goodeve, (University of Sheffield) for her helpful oversight on our manuscript.

Author contributions

KB wrote the manuscript and analyzed all data; KB and NB designed the experiments; NB, KEA, and IMN provided clinical interpretation to results; JD conducted laboratory work; DG and AD supervised.

The authors declare no competing financial interests.

References

1. Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 13: 667-672.
2. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.
3. Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, et al. (2013) Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 34: 1035-1042.
4. Meldrum C, Doyle MA, Tothill RW (2011) Next-generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev 32: 177-195.
5. McCourt CM, McArt DG, Mills K, Catherwood MA, Maxwell P, et al. (2013) Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. PLoS One 8: e69604.
6. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463-5467.
7. Tsiatis AC, Norris-Kirby A, Rich RG, Hafez MJ, Gocke CD, et al. (2010) Comparison of Sanger Sequencing, Pyrosequencing, and Melting Curve Analysis for the Detection of KRAS Mutations: Diagnostic and Clinical Implications. J Mol Diagn 12: 425.
8. Bonetta L (2006) Genome sequencing in the fast lane. Nature Methods 3: 141-147.
9. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, et al. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15: 256-278.
10. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel GD, et al. (2002) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinform 43: 11.10.1-11.10.33.
11. Liu X, Han S, Wang Z, Gelernter J, Yang BZ (2013) Variant callers for next-generation sequencing data: a comparison study. PLoS One 8: e75619.

12. Durtschi J, Margraf RL, Coonrod EM, Mallempati KC, Voelkerding KV (2013) VarBin, a novel method for classifying true and false positive variants in NGS data. BMC Bioinformatics 14 Suppl 13: S2.

13. Reid JG, Carroll A, Veeraraghavan N, Dahdouli M, Sundquist A, et al. (2014) Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics 15: 30.

14. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, et al. (2010) A standardized framework for the validation and verification of clinical molecular genetic tests. Eur J Hum Genet 18: 1276-1288.

15. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

16. Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One 6: e19534.

17. Robasky K, Lewis NE, Church GM (2014) The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 15: 56-62.

18. Mak HC (2011) Next-generation sequence analysis. Nature Biotechnology. 29: 45.

19. Day-Williams AG, Zeggini E (2011) The effect of next-generation sequencing technology on complex trait research. Eur J Clin Invest 41: 561-567.

20. Kilpinen H, Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. Trends Genet 29: 23-30.

21. Keogh MJ, Chinnery PF (2013) Next generation sequencing for neurological diseases: new hope or new hype? Clin Neurol Neurosurg 115: 948-953.

22. Handel AE, Disanto G, Ramagopalan SV (2013) Next-generation sequencing in understanding complex neurological disease. Expert Rev Neurother 13: 215-227.

23. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14: 681-691.

24. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, et al. (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature 486: 400-404.

25. Stingl J, Caldas C (2007) Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. Nat Rev Cancer 7: 791-799.

26. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21: 2213-2223.

27. Roberts ND, Kortschak RD, Parker WT, Schreiber AW, Branford S, et al. (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics 29: 2223-2230.

28. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, et al. (2012) Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 30: 1033-1036.

29. Biesecker LG, Burke W, Kohane I, Plon SE, Zimmern R (2012) Next-generation sequencing in the clinic: are we ready? Nat Rev Genet 13: 818-824.

30. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178-192.

31. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24-26.

32. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

34. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.

36. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80-92.

37. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.

38. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

39. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

40. den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15: 7-12.