## TITLE

# Increased sensitivity of diagnostic mutation detection by re-analysis incorporating local reassembly of sequence reads

## RUNNING HEAD

ABRA reassembly improves test sensitivity

Christopher M. Watson[1,2,3], Nick Camm[1], Laura A. Crinnion[1,2], Samuel Clokie[4], Rachel L. Robinson[1], Julian Adlard[1], Ruth Charlton[1], Alexander F. Markham[2,3], Ian M. Carr[2,3], David T. Bonthron[1,2,3]

1: Yorkshire Regional Genetics Service, St. James's University Hospital, Leeds, LS9 7TF, United Kingdom

2: MRC Single Cell Functional Genomics Centre, University of Leeds, St. James's University Hospital, Leeds, LS9 7TF, United Kingdom

3: MRC Medical Bioinformatics Centre, Leeds Institute for Data Analytics, University of Leeds, Leeds, LS2 9JT, United Kingdom

4: West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust, Birmingham, B15 2TG, United Kingdom

*Corresponding author:
Dr Christopher M. Watson
ORCID ID: 0000-0003-2371-1844
6.2 Clinical Sciences Building
Yorkshire Regional Genetics Service
St James's University Hospital
Leeds, LS9 7TF
United Kingdom

Email: c.m.watson@leeds.ac.uk

Tel: +44 (0) 113 206 5677

Fax: +44 (0) 113 343 8702

## ABSTRACT

### Background

Diagnostic genetic testing programmes based on next generation DNA sequencing have resulted in the accrual of large datasets of targeted raw sequence data. Most diagnostic laboratories process these data through an automated variant-calling pipeline. Validation of the chosen analytical methods typically depends on confirming the detection of known sequence variants. Despite improvements in short-read alignment methods, current pipelines are known to be comparatively poor at detecting large insertion/deletion mutations.

### Methods

We performed clinical validation of a local reassembly tool, ABRA, through retrospective reanalysis of a cohort of more than 2000 hereditary cancer cases.

### Results

ABRA enabled detection of a 96-bp deletion, 4-bp insertion mutation in *PMS2* that had been initially identified using a comparative read-depth approach. We applied an updated pipeline incorporating ABRA to the entire cohort of 2000 cases, and identified one previously undetected pathogenic variant, a 23-bp duplication in *PTEN*. We demonstrate the effect of read length on the ability to detect insertion/deletion variants, by comparing HiSeq2500 (2×101-bp) and NextSeq500 (2×151-bp) sequence data for a range of variants, and thereby show that the limitations of shorter read lengths can be mitigated using appropriate informatics tools.

### Conclusions

This work highlights the need for ongoing development of diagnostic pipelines to maximize test sensitivity. We also draw attention to the large differences in computational infrastructure required to perform day-to-day versus large-scale reprocessing tasks.

**Key points:**

- We demonstrate how reprocessing legacy datasets using improved bioinformatics tools can increase diagnostic test sensitivity and show how variant detection is affected by sequencing read lengths.
- We describe the importance of this approach and highlight the computational infrastructure that is required to undertake large-scale retrospective reanalyses.

## 1. INTRODUCTION

In recent years, massively parallel “next generation” sequencing (NGS) has become the method of choice for molecular genetic screening, succeeding single-locus Sanger sequencing as the gold standard technology. Although most laboratories use Illumina instruments to generate short-read sequencing data, several different target enrichment strategies have been developed and adopted, in which a subset of the genome is selectively enriched before sequencing. One of the most commonly used approaches relies on capture by hybridization; the capture reagent comprises a large number of short (~120-nt) oligonucleotides designed against specific genomic targets (typically the coding regions of chosen genes). Our laboratory was an early adopter of this methodology and we have previously described a custom 36-gene reagent for diagnosis of hereditary cancer [1]. The success of this approach in routine diagnostics has prompted us and others to expand our test portfolios, creating capture reagents aligned to a range of distinct clinical disease groups. More recently, “off-the-shelf” reagents have been made available by commercial manufacturers. As the cost of DNA sequencing continues to fall, and the number of genes that can be concurrently sequenced continues to increase, diagnostic portfolios are likely to converge on small numbers of pre-designed enrichment reagents, paving the way towards more standardized datasets, before, ultimately, whole genome sequencing becomes the single *de facto* approach. Regardless of the approach through which NGS datasets are acquired, genetic diagnostic laboratories are accruing large volumes of targeted sequence data at an unprecedented rate.

The complexity of NGS data processing and interpretation creates a barrier for many laboratories, hindering their adoption of these technologies. As sequencer output has increased, diagnostic laboratories have typically adopted unix-based pipelines that enable automated, efficient and customisable data processing workflows, albeit at the additional cost of specialized bioinformatics expertise and dedicated server infrastructure. In this approach, individual constituent programs are typically obtained from open source repositories; the open access makes it straightforward for new methods to be appraised in a development setting before being incorporated into a production environment. The ability to customise individual pipelines facilitates robust integration of new programs with upstream and downstream informatics processes, as well as optimisation of user-defined parameters when establishing new tests.

While low-coverage whole genome sequencing is an effective method for detecting large, megabase-sized, copy number variants [2], it may also be desirable to detect copy number variants through comparative read-depth analysis of targeted enrichment datasets [3]. While these methods have been successful in identifying exonic deletions and duplications, several factors appear to influence the sensitivity of the approach. These include the GC content or proximity to low complexity sequence of the target genomic region, the total read depth and the number of libraries included in the comparator group, and the enrichment method performed (which may affect the position of the variant within the sequenced DNA fragments). Furthermore, the size and number of adjacent affected exons may also influence the sensitivity of indel detection. Despite the latter limitation, single-exon deletions and duplications are frequently detectable by comparative read-depth analysis of enriched targets; these provide an opportunity to optimise methods that bridge the gap between aligner-based variant callers and approaches based on read depth.

Diagnostic assay validation normally consists of an empirical assessment of an assay's ability to detect variants that are representative of the expected mutation spectrum. The inference is that similar classes of variant should be detected with comparable sensitivity [4]. The robustness of these assumptions is strongest for the most commonly observed variant types (single nucleotide substitutions) and is less robust for heterogeneous indel variants that are harder to detect using short-read NGS technologies. A related problem is that it is often challenging to identify a sufficiently large and representative sample of these "difficult" indel variants with which to perform method validation, due to their rarity in specific disease cohorts [5].

Although such technical factors can be scrutinized objectively, the mutation spectrum underlying many disorders remains incompletely known, due to the paucity of genetic testing to date. Furthermore, traditional molecular genetic methods are struggling to maintain pace with the ever-increasing number of genes that require diagnostic interrogation. This is especially relevant for copy number variant detection, for which the gold standard investigation, at exon-based resolution, has for many years been regarded as a multiplex ligation-dependent probe amplification (MLPA) assay [6]. Only a finite number of targets can be interrogated in a single MLPA reaction, typically fewer than the total number of genes being analysed by a corresponding NGS panel. Consequently, MLPA testing may often impose workload demands that are disproportionate to the overall testing strategy and diagnostic yield. Despite the high sensitivity of MLPA assays, the finding that dosage mutations can be detected using data generated by hybridisation capture enrichment calls into question the long-term utility of this method, especially when NGS datasets are enabling dosage analysis of genes for which MLPA probe-sets do not presently exist [3]. In general, separate informatics methods are required to detect copy number variants, compared to single nucleotide variant / small indels, in NGS data. Furthermore, the limits of resolution of these methods are currently largely non-overlapping, making efforts of the type described herein of considerable practical diagnostic value.

It is important to identify borderline categories of indel variants for further analysis, because they are likely to escape detection in many well-established existing analysis pipelines. Ultimately, sensitivity for detecting such variants is likely to be improved by the advent of longer sequence reads in routine practice. However, for the time being, the wide range of bioinformatics tools and their associated variable parameters, will continue to have a significant effect on pipeline performance. With this in mind, we describe here our implementation of ABRA (assembly-based realigner), a tool capable of performing local reassembly of aligned sequence reads, as an additional component of our diagnostic pipeline [7]. We report an improvement in pipeline sensitivity for large indel variants ranging from a pathogenic deletion/insertion of 96bp/4bp respectively, to a 37bp insertion. We interrogate a cohort of more than 2,000 cases referred for analysis of hereditary cancer genes by reprocessing these data using the refined pipeline and report how ABRA resulted in a marginal gain in test sensitivity for this cohort. Based on our experience, we advocate that routine diagnostic laboratories undertake retrospective reprocessing of existing legacy data, and describe the computational infrastructure implications of such a proactive quality control approach.

## 2. MATERIALS AND METHODS

Patients were referred to the Leeds Genetics Laboratory for diagnostic testing of one or more genes included on either a custom-designed Agilent SureSelect hybridisation enrichment assay or, for one case, the Agilent Focused Exome. Since launching our expanded hereditary cancer service in 2013, the custom reagent has been updated twice, and presently includes probes targeting the exons and immediate flanking sequence of 155 genes causing a range of hereditary cancer disorders (see Supplementary Data File 1 for the genomic coordinates of these regions).

DNA was isolated from peripheral blood samples using either a standard salting out method or the Chemagic™ 360 automated extractor (PerkinElmer, Seer Green, UK). For each sample, approximately 3 μg of genomic DNA was sheared into 200- to 300-bp fragments using a Covaris S2 or E220 (Covaris Inc., Woburn, MA, USA). Fragmented DNA samples were processed through a standard next-generation sequencing library preparation workflow using SureSelect XT reagents (Agilent Technologies, Wokingham, UK). This consisted of end-repair, (A)-addition, adapter ligation and PCR enrichment to create Illumina-compatible paired-end sequencing libraries. Hybridisation capture target enrichment was performed on the whole genome libraries using custom RNA probes, following manufacturer's protocols throughout. The quality of final libraries was confirmed using either an Agilent Bioanalyser or Agilent Tapestation (Agilent Technologies, Wokingham, UK). Post-enrichment final libraries were pooled into batches containing equimolar aliquots of, typically, 16 sequencer-ready samples. Although the workflow was initially performed manually, it has since been automated using a Sciclone G3 liquid handling workstation (PerkinElmer, Seer Green, UK) allowing 96 patient samples to be processed in less than 5 working days. Batches were sequenced either singly on one lane of an Illumina HiSeq2500 rapid-mode flow cell (2 × 101 bp sequencing), or more recently pooled in groups of three and sequenced on an Illumina NextSeq500 (2 × 151 bp sequencing) using a High Output flow cell (Illumina Inc., San Diego, CA, USA). Per-run reagent and sequencer configurations are outlined in Supplementary Data File 2.

Although the post-run data processing pipeline was frequently amended to accommodate software updates and bug fixes, each tool and its position in the workflow remained consistent. Specific details defining the latest version of the pipeline (2016-03-24) are provided. Data processing was performed on a HP DL585G7 4-processor 64-core server upgraded to include 384 GB of RAM and fibre-connected to a 120-TB external storage array (HP Inc., Palo Alto, CA, USA). Sequencer-generated .bcl files were demultiplexed and converted to FASTQ.gz format using bcl2fastq v.2.17.1.14. Adaptor sequences and low-quality bases (Q score ≤10) were identified in per-sample sequencing reads and trimmed using Cutadapt v.19.1 (https://github.com/marcelm/cutadapt) [8]. FastQC v.0.11.5 was run to confirm sequencing performance on alignment-ready FASTQ.gz files (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequencing reads were aligned to an indexed human reference genome (hg19) using BWA MEM v.0.7.13 (http://bio-bwa.sourceforge.net) [9]. Reads were sorted by chromosome coordinate and PCR duplicates were marked using Picard v.2.1.1 (http://broadinstitute.github.io/picard/) creating a 'processed.bam' file. The Genome Analysis Toolkit (GATK) v.2.3-4Lite was used to perform indel realignment and base quality score recalibration on the processed.bam file, following the GATK best practice workflow (http://gatkforums.broadinstitute.org/gatk) [10]. Variant calling was performed using the GATK UnifiedGenotyper with the minimum indel fraction argument adjusted to 0.01. Detected variants were exported in variant call format (VCF) and this file was annotated with biologically meaningful information using Alamut Batch standalone v.1.4.4 (database v.2016.03.04) (Interactive Biosoftware, Rouen, France). The pathogenicity status of identified variants was assessed using a set of custom-designed Microsoft Excel spreadsheets and interpretation was performed according to ACGS best practice guidelines [11]. Clinical reports were generated for all patients based on the gene requested for analysis. Coverage metrics were calculated using the GATK walkers DepthOfCoverage, CallableLoci and CountReads to assess the quality of the interrogated genomic loci. Aligned sequence reads were visualised with the Integrative Genome Viewer v.2.3.80 (http://software.broadinstitute.org/software/igv/) [12]. To identify dosage variants, a custom-designed method relying on an intra-batch comparative read depth analysis was performed. The effect of ABRA v.0.97 was assessed on processed.bam files (https://github.com/mozack/abra) [7].

Comprehensive reprocessing of the legacy dataset was performed using the Leeds MRC Medical Bioinformatics Centre high-performance computer MARC1, which consists of 57 HP BL460 blades, each comprising a dual socket 10-core Intel E5-2660v3(2.6GHz) processor and 256GB of DDR4 2133MHz RAM (http://arc.leeds.ac.uk/systems/marc1/). Pipeline flow control was reconfigured to work on a per-sample, rather than per-batch, basis. Pipeline components that were updated included Picard (from v.2.1.1 to v.2.4.1), Alamut Batch standalone (from v.1.4.4 to v.1.5.1 and database from v.2016.03.04 to v.2016.06.16) and GATK (from v.2.3-4Lite to v.3.6-0). The latter enabled use of the HaplotypeCaller (which was implemented with the additional argument '--dontTrimActiveRegions'). Two additional variant-calling algorithms were also incorporated into the development pipeline; Platypus v.0.8.1 (http://www.well.ox.ac.uk/platypus) [13] and Varscan v.2.4.2 (http://dkoboldt.github.io/varscan/) [14]. Annotated variants were filtered using a custom AWK script that interrogated genes requested at the point of referral and retained only those variants whose rsClinicalSignificance field was 'pathogenic', codingEffect field was listed as 'in-frame, frameshift or stop gain' or whose variant location was at the ±1 invariant splice site.

Sanger sequencing was used to confirm all described variants; manufacturer's protocols were followed throughout (Applied Biosystems, Paisley, UK). Primer details and thermocycling conditions are available on request. Sequence chromatograms were analysed using Mutation Surveyor v.3.2 (SoftGenetics LLC, State College, PA, USA) and Chromas v.2.6.2 (http://technelysium.com.au/wp/chromas/).

## 3. RESULTS

We have previously implemented an NGS pipeline for detecting variants from hybridisation enrichment assays [1]. This uses the GATK UnifiedGenotyper for the identification of single nucleotide and small indel variants, together with a custom-designed comparative read-depth method for detecting exon-sized deletions and duplications. The detection limits of these complementary approaches are exemplified by the heterozygous *PMS2* variant c.24-12_107delinsAAAT (NM_000535.6), which was detected using the comparative read-depth method, but not by the GATK UnifiedGenotyper. To examine the effect of read alignment accuracy on the sensitivity of the UnifiedGenotyper pipeline, we deployed ABRA, a *de novo* reassembly tool [7], and assessed its performance using sequence data from this case. Analysis of the ABRA-processed BAM file revealed that the total number of reads mapping to the variant locus had increased, as had the proportion of deletion-spanning gapped alignments, which rose from 0% to 23% (Table 1). Evaluation of the VCF file generated from the ABRA-processed BAM file revealed that although the variant had been detected by the UnifiedGenotyper, the entry was split across three output rows, thus preventing Alamut Batch from outputting the correct Human Genome Variation Society (HGVS) nomenclature for this variant.

Having observed that ABRA improved the proportion of gapped read alignments towards the expected 50:50 allelic ratio, we investigated ABRA's effect on three further deletion-containing variants of varying size; 9 bp, 31 bp and 40 bp respectively. At each locus, we detected an increase in both the total number and proportion of gapped reads (Table 1). Visualisation of the read-alignment coverage profiles at these loci was consistent with the reported number of gapped reads having increased (Figure 1).

To investigate the effect of longer sequence reads on these metrics, the libraries were resequenced using a NextSeq500, which generated paired-end 151-bp data. For each of these libraries, the proportion of natively mapped gapped alignments (*i.e.* alignments generated without the use of ABRA) was greater than was observed for corresponding 101-bp datasets (Table 2). The final proportion of gapped alignments, after ABRA reassembly, was comparable to that seen for the 101-bp sequencing datasets. When the *PMS2* c.24-12_107delinsAAAT variant-containing library was re-sequenced, no gapped alignments were natively mapped *(i.e.* in the absence of ABRA), a result identical to that obtained for the 101-bp dataset.

Our retrospective cohort of previously analysed hereditary cancer referrals comprised 2,042 libraries, pooled into 131 batches. Of these, 116 (89%) batches were sequenced using a HiSeq2500 and the remaining 15 (11%) using a NextSeq500. Data processing run times were correlated to the per-batch library for which the largest number of total reads were generated (Supplementary Figure 1). As NextSeq500 read lengths are 50% longer, the corresponding compute time required to complete the data processing workflow increased accordingly. The run time required to reprocess the entire legacy dataset, using our existing hardware and pipeline infrastructure, was estimated at 26 days. We therefore used the Leeds MRC Medical Bioinformatics Centre high-performance compute cluster MARC1 to implement an updated development pipeline, which included both ABRA and three separate variant-calling tools (the GATK HaplotypeCaller, Platypus and Varscan). A matrix of detection sensitivity for each variant calling tool is displayed, for each reported variant, in Supplementary Table 1.

The entire legacy dataset was processed in four scheduled batches of approximately 500 cases, and took less than 3 days to run.

In this retrospective analysis, we identified a single case for which a pathogenic mutation had not been detected using our existing diagnostic pipeline. This variant, a heterozygous 23-bp duplication in *PTEN*, is located at position c.342_364dup (NM_000314.4) p.(Ile122Lysfs*20) and is predicted to cause a frameshift in the translated protein. The gene affected by this variant is consistent with the patient's phenotype of macrocephaly, and thyroid and ovarian cancers, in addition to a further family history of cancer. Although the UnifiedGenotyper did not detect the variant from the original processed.bam file, visual inspection of the alignment profile, which was generated from paired-end 101-bp reads, revealed an abnormal coverage profile (Figure 2A). Reprocessing the sequence data for this case, using the ABRA-incorporated pipeline, enabled the UnifiedGenotyper to detect this variant. The read alignment coverage plot was altered following ABRA reprocessing, demonstrating the effect of *de novo* read assembly at this locus (Figure 2B). The number of reads with an 'insertion' at the duplication site (chr10:89692857) increased from 124 before ABRA reassembly, to 2,316 following ABRA reassembly.

Retrospective examination of indel variants reported by the laboratory revealed a 24-bp duplication, located at position c.9_32dup p.(Ala4_Pro11dup) (NM_000077.4) in *CDKN2A*, that had been previously detected in our cohort of processed patients. Although this variant is 1 bp longer than the *PTEN* duplication c.342_364dup (that was originally missed), the *CDKN2A* patient data had been generated using the NextSeq500, with 151 bp rather than 101 bp read lengths. To assess whether it was the increased read length or the local sequence context, that enabled the variant to be detected using our normal diagnostic pipeline, we resequenced the original *CDKN2A* c.9_32dup variant-containing library. The variant was detected in the paired-end 101-bp dataset when ABRA reassembly was performed, but was not detectable in the absence of ABRA. To further assess the effect of increased read length, we re-sequenced the *PTEN* c.342_364dup variant-containing library, this time with 151-bp (as opposed to 101-bp) reads, and noted that the variant was now detected using our standard pipeline that did not include ABRA. The proportion of pre- and post-ABRA gapped/insertion-containing read alignments, for each variant, are reported in Supplementary Table 2.

In summary, this work indicated that without realignment by ABRA, 151-bp reads were sufficient to detect the described *PTEN* 23-bp and *CDKN2A* 24-bp duplications. In contrast, we were unable to detect a 37-bp insertion in *SLC26A4* (that was identified following manual scrutiny of the aligned reads in the IGV). This case had been referred for diagnostic genetic testing of deafness-associated genes. It was the initial identification of the heterozygous *SLC26A4* point mutation c.716T>A (NM_000441.1) p.(Val230Asp) (a known recessive pathogenic variant) that prompted further scrutiny of the sequencing data. Reprocessing these data through the ABRA-incorporated pipeline enabled the UnifiedGenotyper to detect the insertion (c.1416_1417ins37), increasing the number of insertion-containing reads from 93 (total count 498) to 205 (total count 497) before and after ABRA processing, respectively.

**4. DISCUSSION**Our experience indicates that ABRA, a *de novo* reassembly algorithm that remaps reads with greater accuracy than is initially achieved, can contribute to an improved ability to detect indel mutations in this important “no-man’s land”. In a series of variants that escaped detection, either in current or legacy short-read data sets, ABRA reassembly improved read alignment metrics at variant loci towards expected biological values (a 50:50 allelic ratio). This has prompted us to incorporate ABRA into our revised diagnostic pipeline, and indicates that we have at least partly bridged the gap referred to above.

Whenever an improvement in diagnostic sensitivity occurs, difficult decisions regarding the re-analysis of existing patient cohorts must be addressed. Diagnostic laboratories rarely, if ever, reanalyse legacy data, because most of their data processing resources are typically consumed by ongoing operational requirements. In our case, more than 2,000 inherited cancer cases clearly merited re-analysis. However, our local computational infrastructure was insufficient to undertake such a task, which would have blocked the server used for diagnostic analysis for several weeks. In contrast, by using a high-performance research computing cluster, we completed the reprocessing task in less than 3 days. An alternative approach would have been to devote a proportion of our existing diagnostic infrastructure to retrospective reprocessing tasks. However, the rapid pace of change in software and instrumentation may ultimately make such a long-term, low-intensity approaches impractical.

That technical reassessments of raw sequencing data are rarely undertaken is perhaps surprising, given the low concordance rates reported between different combinations of alignment and variant-calling algorithms [15]. For many years, diagnostic laboratories have addressed issues of technical accuracy through external quality assurance programs, by inter-laboratory sample exchanges or by obtaining specimens from independent suppliers such as the National Institute for Biological Standards and Control (http://www.nibsc.org).

Recognising the limited number of genotypes that can be tested by these approaches, the Genome in a Bottle Consortium was established to provide comprehensive authoritative characterisation of whole human genome samples for technology development and optimisation purposes (http://jimb.stanford.edu/giab/). However, given that characterisation of these materials is ongoing (and that new enrichment and sequencing technologies are constantly being developed), it seems unlikely that clinical diagnostic pipelines will come to be regarded as 'static' or 'finished' in the near future.

Complementary benchmarking tools are emerging to assist laboratories with in-house validation of standard reference materials (https://github.com/ga4gh/benchmarking-tools/). Despite these, the number of programs involved in a laboratory's data processing pipeline often necessitates a pragmatic approach, resulting in scheduled updates, rather than the immediate implementation of new software versions following their release. From a diagnostic perspective, national accreditation bodies and local validation procedures rightly mandate verification of modified pipelines using data from previously analysed samples [4]. As continued experience is gained, it is likely that the refinement of existing "best-practice" guidelines and version control methodologies will help to ease the burden of this complex task and often resource intensive task [cite http://www.acgs.uk.com/media/1025075/ngs_bioinformatics_bpg_final_version_2016.pdf]

Given that we initially assessed ABRA for its improved performance with deletion mutations, it is noteworthy that the first “missed” mutation we identified was a 23-bp *PTEN* duplication. Comparative read-depth approaches work well for detecting deletions, because the “soft-clipped” reads show up as changes in per-base read depth. In contrast, intra-read insertions generally fail to increase the per-base read-depth profile. This exemplifies the difficulties in defining limits of detection across heterogeneous mutation sets, without large numbers of testable variants. Our examples of 23-bp (*PTEN*), 24-bp (*CDKN2A*) and 37-bp (*SLC26A4*) pathogenic insertions all conform to a trend in which longer reads and local reassembly (ABRA) both enhance mutation detection. We focus on the value of ABRA because it permits the retrospective reanalysis of clinical datasets without expensive re-sequencing.

There are additional potential benefits from the wholesale re-analysis of diagnostic NGS datasets. These include the ability to adapt to changes in external data sources, including the reference genome used for read alignment. When diagnostic laboratories make the (infrequent) transition to a new genome build, interoperability between the genomic coordinates of legacy datasets and those of future referrals is likely to require retrospective re-analysis.

Here we highlight and advocate a change to established genetic diagnostic practice. We show that it is possible for a small regional diagnostic laboratory to engage local research expertise and facilities, and accomplish the reprocessing of accumulated diagnostic genomic datasets.

Our recommendation to implement retrospective analysis of diagnostic cohorts has significant resource implications, made clear by our need to use a confidential research computing platform and to spend time porting the workflows between different secure server infrastructures. Nonetheless, the overall cost was small, compared to the investment required to establish NGS-based diagnostics. We also recognize that the current fashion for centralized national genomics infrastructures [16-18] may adversely affect local investment in such quality assurance activities. Despite this, it seems likely in future, given that laboratories require robust validation procedures in order to obtain clinical accreditation, that analyses comparable to those we present here will be required as one part of quality assurance.

## Compliance with Ethical Standards

### Conflict of interest

All authors (CMW, NC, LAC, SC, RLR, JA, RC, AFM, IMC, DTB) declare that they have no competing interests.

### Funding

This work was supported by grants MR/M009084/1 and MR/L01629X/1 awarded by the UK Medical Research Council.

### Ethical approval & informed consent

Written informed consent was obtained from all reported individuals. Ethical approval was granted by the Leeds East Research Ethics Committee (07/H1306/113).

## 5. REFERENCES

1. Watson CM, Crinnion LA, Morgan JE, Harrison SM, Diggle CP, Adlard J, Lindsay HA, Camm N, Charlton R, Sheridan E, Bonthron DT, Taylor GR, Carr IM. Robust diagnostic genetic testing using solution capture enrichment and a novel variant-filtering interface. Hum Mutat. 2014;35:434-441. doi:10.1002/humu.22490

2. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, McKinley C, Egan P, Ross L, Hayward B, Morgan J, Davidson L, MacLennan K, Ong TK, Papagiannopoulos K, Cook I, Adams DJ, Taylor GR, Rabbitts P. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. 2010;38:e151. doi:10.1093/nar/gkq510

3. Watson CM, Crinnion LA, Berry IR, Harrison SM, Lascelles C, Antanaviciute A, Charlton RS, Dobbie A, Carr IM, Bonthron DT. Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping. BMC Med Genet. 2016;17:1. doi:10.1186/s12881-015-0265-z

4. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V, Wallace A; EuroGentest Validation Group. A standardized framework for the validation and verification of clinical molecular genetic tests. Eur J Hum Genet. 2010;18:1276-1288. doi:10.1038/ejhg.2010.101

5. Deans Z, Watson CM, Charlton R, Ellard S, Wallis Y, Mattocks C, Abbs S. Practice guidelines for Targeted Next Generation Sequencing Analysis and Interpretation. Association for Clinical Genetic Science. 2015. http://www.acgs.uk.com/media/983872/bpg_for_targeted_next_generation_sequencing_-_approved_dec_2015.pdf. Accessed 7 Jul 2017.

6. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 2002;30:e57.

7. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics 2014;30:2813-2815. doi:10.1093/bioinformatics/btu376

8. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10-12. doi:10.14806/ej.17.1.200

9. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-1760. doi:10.1093/bioinformatics/btp324

10. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for

variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491-498. doi:10.1038/ng.806

11. Wallis Y, Payne S, McAnulty C, Bodmer D, Sistermans E, Robertson K, Moore D, Abbs S, Deans Z, Devereau A. Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics. Association for Clinical Genetic Science. 2013. http://www.acgs.uk.com/media/774853/evaluation_and_reporting_of_sequence_variants_bpgs_june_2013_-_finalpdf.pdf. Accessed 7 Jul 2017.

12. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178-192. doi:10.1093/bib/bbs017

13. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, WGS500 Consortium, Wilkie AO, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912-918. doi:10.1038/ng.3036

14. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-576. doi:10.1101/gr.129684.111

15. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875. doi:10.1038/srep17875

16. Caulfield M, Ainsworth C. Q&A: Mark Caulfield. National genomics. Nature 2015;527:S5. doi:10.1038/527S5a

17. Project Team SG. The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease. IEEE Pulse. 2015;6:22-26. doi:10.1109/MPUL.2015.2476541

18. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K: Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47:435-444. doi: 10.1038/ng.3247

**Figure 1:** Alignment read coverage plots showing the effect of ABRA processing at variant loci. The total number of reads flanking each variant site, and the number of gapped alignments, is increased for each sample following ABRA reassembly. This is visible from the increased height of the profiles at the margins of the variant "trough". Reads in the trough represent the invariant allele. The y-axis is scaled to show a maximum value of 2,000×, allowing plots to be comparable between samples. Scale bars denote a 20-bp interval. Variants are annotated according to the following transcripts: *BRCA2* (NM000059.3), *MLH1* (NM_000249.2), *BRCA1* (NM_007294.3) and *PMS2* (NM_000535.6). The libraries were prepared manually using hybridisation capture reagent version 2 and sequenced on a HiSeq2500.

**Figure 2:** Alignment read depth plots displaying the *PTEN* c.342_364dup (NM_000314.4) variant locus for sequence data processed using **(A)** the initial diagnostic pipeline and **(B)** the ABRA-incorporated pipeline. The y-axis is set to a maximum value of 7000× and is therefore comparable between panels. Scale bars denote a 20-bp interval. The library was prepared manually using hybridisation capture reagent version 1 and sequenced on a HiSeq2500.

**Supplementary Figure 1:** Per-batch pipeline run times are correlated to the library with the greatest number of reads in each batch. Data generated using a HiSeq2500 is processed more quickly than that generated using a NextSeq500, due to the shorter read length.

**Table 1: The proportion of gapped read alignments for paired-end 101-bp reads before and after ABRA reassembly.**

| **Sample ID** | **1** | | **2** | | **3** | | **4** | |
|---|---|---|---|---|---|---|---|---|
| **Gene** | *BRCA2* | | *MLH1* | | *BRCA1* | | *PMS2* | |
| **Transcript** | NM_000059.3 | | NM_000249.2 | | NM_007294.3 | | NM_000535.6 | |
| **c.Nomen** | c.8736_8744del | | c.197_207+20del | | c.1175_1214del | | c.24-12_107delinsAAAT | |
| **Sequence affected** | 9 bp deleted | | 31 bp deleted | | 40 bp deleted | | 96 bp deleted and 4 bp inserted | |
| | - | + | - | + | - | + | - | + |
| **Number of 5' flanking reads** | 1433 | 1546 | 1271 | 1440 | 1409 | 1714 | 1092 | 1151 |
| **Number of 3' flanking reads** | 1405 | 1510 | 1163 | 1366 | 1395 | 1716 | 585 | 810 |
| **Mean number of flanking reads** | 1419 | 1528 | 1217 | 1403 | 1402 | 1715 | 839 | 981 |
| **Number of gapped alignments** | 438 | 656 | 193 | 565 | 142 | 767 | 0 | 225 |
| **Gapped reads as a proportion of total reads at the variant site (%)** | **31** | **43** | **16** | **40** | **10** | **45** | **0** | **23** |

BAM file read metrics before (-) and after (+) ABRA reassembly was performed. Libraries were prepared manually using hybridisation capture reagent version 2 and sequenced using a HiSeq2500.

**Table 2: The proportion of gapped read alignments for paired-end 151-bp reads before and after ABRA reassembly.**

| **Sample ID** | **1** | | **2** | | **3** | | **4** | |
|---|---|---|---|---|---|---|---|---|
| **Gene** | *BRCA2* | | *MLH1* | | *BRCA1* | | *PMS2* | |
| **Transcript** | NM_000059.3 | | NM_000249.2 | | NM_007294.3 | | NM_000535.6 | |
| **c.Nomen** | c.8736_8744del | | c.197_207+20del | | c.1175_1214del | | c.24-12_107delinsAAAT | |
| **Sequence affected** | 9 bp deleted | | 31 bp deleted | | 40 bp deleted | | 96 bp deleted and 4 bp inserted | |
| | - | + | - | + | - | + | - | + |
| **Number of 5' flanking reads** | 2538 | 2625 | 2779 | 2994 | 2719 | 3067 | 1680 | 1855 |
| **Number of 3' flanking reads** | 2454 | 2560 | 2585 | 2886 | 2704 | 3062 | 907 | 1362 |
| **Mean number of flanking reads** | 2496 | 2593 | 2682 | 2940 | 2712 | 3065 | 1294 | 1609 |
| **Number of gapped alignments** | 858 | 1049 | 680 | 1192 | 609 | 1310 | 0 | 455 |
| **Gapped reads as a proportion of total reads at the variant site (%)** | **34** | **40** | **25** | **41** | **22** | **43** | **0** | **28** |

BAM file read metrics before (-) and after (+) ABRA reassembly was performed. Libraries were prepared manually using hybridisation capture reagent version 2 and sequenced using a NextSeq500.