Accurate sample assignment in a multiplexed, ultra-sensitive, high-throughput sequencing assay for minimal residual disease

Jack Bartram[1,2], Edward Mountjoy[3], Tony Brooks[4], Jeremy Hancock[5], Helen Williamson[5], Gary Wright[2], John Moppett[6], Nick Goulden[2], Mike Hubank[1,4*]

[1]Genetics and Genomic Medicine Programme, Institute of Child Health, University College London, UK; [2]Depatment of Haematology, Great Ormond Street Hospital for Children, London, UK; [3]School of Social and Community Medicine, University of Bristol, UK; [4]UCL Genomics, Institute of Child Health, University College London, UK; [5]Bristol Genetics Laboratory, Southmead Hospital, North Bristol NHS Trust, UK; [6]Department of Paediatric Haematology/Oncology, Royal Hospital for Children, Bristol, UK

Short running head: Accurate multiplexed high-throughput MRD

[*]**Correspondence**: Dr Mike Hubank**,** Genetics and Genomic Medicine Programme, Institute of Child Health, University College London, 30 Guilford Street, London, WC1N 1EH, Telephone: (+44) 02079052266, Fax: Email: m.hubank@ucl.ac.uk

Text pages: 16

Text word count: 4388

Abstract word count: 192

Figures: 5

Tables: 5

Reference count: 46

Grant numbers: GOSHCC W1069; Above & Beyond project 4/2012-13

Conflicts of interest: The authors have no conflicts of interest to declare

ABSTRACT

High throughput sequencing (HTS) (next generation sequencing) of the rearranged immunoglobulin and T-cell receptor genes promises to be cheaper and more sensitive than current methods for monitoring minimal residual disease (MRD) in patients with acute lymphoblastic leukemia. However, adoption of new approaches by clinical laboratories requires careful evaluation of all potential sources of error and the development of strategies to ensure the highest accuracy. Timely and efficient clinical use of HTS platforms will depend on combining multiple samples (multiplexing) in each sequencing run. Here we examine *immunoglobulin heavy chain* gene HTS on the Illumina MiSeq platform for MRD (HTS-MRD). We identify errors associated with multiplexing that could potentially impact on the accuracy of MRD analysis. We optimise a strategy combining high purity, sequence-optimised oligonucleotides, dual-indexing and an error-aware demultiplexing approach to minimise errors and maximise sensitivity. We present a probability-based demultiplexing pipeline, Error-Aware Demultiplexer (EAD) - that is suitable for all MiSeq sequencing strategies and accurately assigns samples to the correct identifier without excessive loss of data. Finally using controls quantified by digital PCR, we show that HTS-MRD can accurately detect as few as 1 in 10^6 copies of specific leukemic MRD.

Introduction

The accurate determination of minimal residual disease (MRD) during the early months of therapy in acute lymphoblastic leukemia (ALL), particularly childhood ALL, is well established as a biomarker for guiding therapy[1-6]. Current methods for MRD measurement – allele-specific real-time quantitative (ASO-RQ) PCR of clone-defining immunoglobin (IG)/T-cell receptor (TCR) gene rearrangements in the patients' leukemia[7, 8] and flow cytometric (FC) tracking of leukemia associated immunophenotypes[1, 9] - are both expensive, time consuming and suffer from technical limitations. ASO-RQ PCR requires assays tailored to each individual patient and, depending on template availability and primer selection has a maximal sensitivity of $1{:}1x10^{-5}$ due to non-specific background amplification[10]. This prevents the identification of even lower risk patients who could benefit from safer protocols that further reduce treatment-related mortality. Standardisation of FC is difficult, requires experienced operators (especially pediatric samples), and inter-operator variation can lead to inconsistent reporting[11]. Finally, clonal architecture is dynamic. When disease relapse occurs, it can involve clones that were not identified, or only viewed as minor clones, at diagnosis and therefore were not tracked[12].

Advances in high throughput sequencing (HTS) offer a potential solution to these problems. Highly parallel HTS can be employed to sequence the rearranged *VDJ* (*Variable, Diverse* and *Joining*) of the *immunoglobulin heavy chain* (*IGH*) genes, which encode the hypervariable CDR3 domain. Combined with the high capacity of HTS, this allows a single, clone-unbiased, and highly sensitive test to be applied to all patients, revealing persisting or evolving clones, potentially even if these were not the defining clones at presentation. Importantly, HTS generates exact nucleotide sequences for all clones which are unique to each leukemic clone can be traced through subsequent follow-up analysis. Several reports have demonstrated the potential advantages of HTS for the molecular characterization of haematological malignancies[13-21].

In order to translate these advances, it is necessary to establish a high throughput sequencing MRD (HTS-MRD) workflow that is practical to operate in a clinical laboratory, cost effective, and

demonstrated to be rapid, accurate and reproducible[22]. Continuing development of HTS technologies has resulted in cheaper platforms, such as the MiSeq (Illumina), that operate at a capacity which more appropriately matches the demands of turnaround time and cost required in clinical practice[23-25]. Timely and efficient clinical use of HTS platforms depends on combining multiple samples in each sequencing run, or multiplexing. Indexing of samples with unique "index sequences" allows a wide degree of multiplexing to be achieved per run, maximising use of sequencing space. However the technical limitations of this multiplexing strategy require detailed investigation – in particular, how accurately are sequences matched with patient on the basis of the indexing? Small errors in MRD assignment could have serious clinical implications. Incorrectly stratifying a patient could result in under or over treating with exposure to unnecessary toxicity or increased risk of relapse. HTS-MRD experiments generate large datasets. Interpretation and reproducibility of these data are of crucial importance in the development of a reliable clinical assay[26]. Use of these technologies in clinical practice has been cautioned until they are fully validated[27].

The potential for primer bias influencing MRD detection has been extensively modelled by other groups[28], so we do not address this here. The amplification strategy used for the current ASO-RQ PCR assay is already clinically approved[10] and therefore provides the most practical basis for translation into a HTS-MRD assay. After evaluating alternatives, we chose the MiSeq as a suitable platform for delivery as it is able to sequence single read (unidirectional) libraries of sufficient length (minimum 150 nucleotides) and quality to identify clones in less than 24 hours. MiSeqDx has been FDA approved for diagnostic use in cystic fibrosis[29].

Sequencing multiple indexed samples per run reduces costs and increases scalability; however it introduces the potential risk of assigning reads to the wrong patient sample (misassignment). We therefore systematically investigated potential sources of multiplexing error on the MiSeq that could reduce the sensitivity and accuracy of MRD identification. These include index cross-contamination, sequencing error, misassignment of indices, run-to-run carry over, and the accuracy of the demultiplexing algorithm.

We found significant problems with "off-the-shelf" solutions and workflows for multiplexed amplicon sequencing, with low but unacceptable levels of sample misassignment, which could potentially lead to false-positive calls in clinical use. We overcame these issues by applying a dual-indexing strategy similar to that described by Kircher *et al*.[30], using high quality preparations of index oligonucleotides[31], and by developing an informatics pipeline to filter out low quality sequencing reads and reduced quality index reads. Finally, we implement our workflow and demonstrate an accurate quantification strategy using a reference "spike-in" method quantified by digital PCR (dPCR) that potentially exceeds the accuracy of current approaches by at least ten-fold.

Materials and Methods

Samples and cell lines

Ethical approval was given (Research Ethics Committee reference 13/LO/1262) for use of appropriately consented material from patients with B-lineage ALL at Great Ormond Street Hospital for Children. Forty one pre-treatment and eight post-induction chemotherapy bone marrow (BM) samples were obtained. Pooled "normal" lymphocyte DNA came from the UK National Blood Service. The leukemic cell lines SUPB15, REH and TOM-1 were from DSMZ and BEL-1 was kindly donated by Dr RW Stam (Rotterdam, NL).

Sample Preparation

The mononuclear cell fraction of BM samples was isolated following centrifugation on Ficoll-Hypaque (density 1.077g/l). Authentication of cell lines was performed by short tandem repeat analysis using the PowerPlex-16 system (Promega). DNA was extracted according to standardised protocols[32] using QIAamp DNA MiniKit (Qiagen). DNA concentration was estimated using spectrophotometry (Nanodrop, Thermo Scientific), then accurately quantitated by RQ-PCR using albumin as a control/reference gene.

Oligonucleotide synthesis

Single-index strategy indices were HPLC purified and synthesised (Sigma-Aldrich) as per Kozarewa *et al.*[33] We designed new dual-indices (Table 1) synthesised using the TruGrade process (Integrated DNA Technologies)[31].

Digital PCR

The CDR3-encoding regions of *IGH* genes of the cell lines described above were amplified and sequenced (HTS and Sanger) to ensure clonality and purity of sequence (Supplementary figures 1 and 2). TaqMan assays were then designed for the unique CDR3 region. Reactions containing TaqMan Gene Expression Master Mix (ABI), GE sample loading reagent (Fluidigm), TaqMan assay and template DNA, were pipetted into the loading inlets of a 12.765 Digital Array (Fluidigm). The BioMark IFC controller MX (Fluidigm) was used to uniformly partition the reactions into the panels. Template molecules are partitioned throughout the panels with a high degree of randomness and independence[34]. Absolute copy number quantification of cell line "spike-in" *IGH* CDR3-encoding regions was performed by dPCR, using the BioMark Real-Time PCR System (Fluidigm). For each 12.765 dPCR array template DNA was analysed in triplicate.

HTS strategy

IGH genes were amplified by multiplex PCR using AmpliTaq Gold (ABI) in a 2-stage PCR (Figure 1 and Supplementary table 1). In the first stage, *IGH* family primers were modified to contain partial MiSeq adaptor sequences. First stage products were purified using solid phase reversible immobilisation (SPRI) beads (Agencourt AMPure XP, Beckman Coulter), before fluorometer quantification (Qubit, Invitrogen) and DNA Bioanalyser (Agilent) quality assessment. The purified product formed the template in a second stage PCR using NEBNext High Fidelity master mix (New England Biolabs) in which sample specific indices and full MiSeq adaptor sequences were added. The indexed samples were again purified, quantified, and then normalised to create the sequencing library pool. Sequencing libraries were re-quantified by RQ-PCR using the KAPA library quantification kit for Illumina sequencing (Kapa Biosystems) or TaqMan Gene Expression Assay for Illumina Library

Quantification (Life Technologies). Accurate quantification of molecules bearing appropriately ligated Illumina adaptors was crucial to ensure optimal cluster density for sequencing. Sequencing mix with 5-10% PhiX (to offset low early cycle sequence complexity) was loaded onto the MiSeq following the Illumina protocol "Preparing Libraries for Sequencing on the MiSeq" (Illumina Inc., San Diego, CA). We used a single-end read from *J* to *V* to ensure optimal quality over the CDR3-encoding region. Indexing reads were performed to identify the 8 base pair (bp) index sequences (single or dual-indexed). Sequencing runs performed in this study are listed in Table 2. The pre-processed (i.e. multiplexed bcl) data files discussed are available in controlled access format in the European Genome-phenome archive (EGAS00001001303). Custom bioinformatics pipelines were used to identify, cluster and annotate sequencing reads.

For evaluation of our workflow in clinical practice we used 5 randomly chosen real patient samples (MRD quantified for patient stratification by ASO RQ-PCR as gold standard). HTS workflows with and without the technical safeguards described in the paper, where then followed. Input sample was 100 000 cell equivalent DNA. The HTS preparations also included separately indexed diagnostic sample of each patient at a one in ten dilution to further test the versatility of the workflow. MRD samples were "spiked" with 1, 10 and 100 copies of SUPB15, TOM-1, and REH *IGH* sequences for quantification purposes. To assess the potential limit of detection of HTS-MRD we also spiked the same quantity of cell line reference DNA into one million cell equivalents of pooled donor lymphocyte DNA.

Results

Misassignment of indices

We first examined the accuracy with which oligonucleotide indexing assigned reads to the correct sample using standard laboratory and analytical methods. Initially we tagged each sample with a single P7 index composed of 8-mer oligonucleotides chosen at random a panel of 96 described by Kozarewa *et al.*[33]. Sequencing runs were initially demultiplexed using the on-board MiSeq software which bins samples according to the 8-mer index. In addition to indices used for sample tagging in the

experiment, we also instructed the demultiplexing program to search for the remaining indices that had been synthesised, but were not included in the experiment.

In a typical experiment (A7BK7, Table 2), with total reads passing filter 18.59 million, 89.1% (16 538 154) reads were assigned to indices corresponding to samples included in the experiment (Supplementary table 2), with 10.9% (2 036 018) undetermined reads not assigned to indices after demultiplexing (Supplementary table 3). Of the undetermined reads, 55% aligned to PhiX genome, added for quality control purposes and to improve cluster resolution, 41.7% to rearranged *IGH* reads; 0.002% to non-rearranged *IGH* reads and 3.3% to other non-*IGH* reads, aligned elsewhere on the genome (Supplementary table 3, Supplementary figure 3). We found that 0.12% - 23,044 reads were misassigned to one of the 68 indices not used to tag samples in the run (Figure 2, Supplementary table 3). Overall this suggested that > 1 in 1000 reads are misassigned by the standard on-board MiSeq demultiplexing pipeline. This could be caused by sequencing error or factory oligonucleotide cross-contamination.

Quality scores of misassigned reads

To assess the extent of sequencing error we generated quality scores for the misassigned reads and their associated index reads using FASTQC - www.bioinformatics.babraham.ac.uk/projects/fastqc/ - (representative examples from run A7ELC on Table 2 are shown in Figure 3). We discovered that the quality of misassigned sample and index reads (Figure 3D, F) is inferior to that of reads assigned to real samples (Figure 3A, C) and deteriorates with sequence length, resulting in poor mean quality scores (Figure 3E). For the index reads, average Phred quality score for correctly assigned reads was 32.5 (range 30.4 – 35.2) (Figure 3C) compared to 19.3 (range 18.4 – 22.1) for misassigned index reads (Figure 3F). This gives a mean difference of 13.3 which was highly significant with $p < 0.00001$ (95% CI 11.5 – 15.2). These results indicate that misassignment stems at least in part from poor quality index reads and read quality filters are required to ensure the most accurate demultiplexing strategy.

Redesigned oligonucleotide indices

Misassignment may also result from oligonucleotide cross-contamination during synthesis, as previously shown by *Quail et al*[31]. We designed a new set of 8-mer oligonucleotide indices with high Hamming distance to optimise maximal sequence difference, homopolymer length and GC/AT balance[35, 36]. These were then synthesised using the high purity TruGrade process which reduces the risk of factory cross-contamination of oligonucleotide stocks[31]. We adopted a dual-indexing approach previously shown to improve accurate demultiplexing[30], increasing total index nucleotides to 16 bp (an 8 bp index on either end of the amplicon). Twenty-four i7 and sixteen i5 indices were designed (Table 1). In order to monitor frameshift errors, some oligonucleotides were designed with sequences which maintained a high Hamming distance compared to other sequences, but which were shifted by a single base position (Supplementary table 4). Initially we synthesised only indices i7 1-12 and i5 1-8.

A sequencing run was performed using this new indexing strategy (A7ELC, Table 2). Thirty-one samples were multiplexed and sequenced. All index combinations of the indices i7 1-12 and i5 1-8 were entered on the sample sheet, giving a total comprised of 31 double-indexed samples, 65 index combinations synthesised but not used in the experiment, and 288 index combinations where oligonucleotides were not synthesised at all.

We found that for the majority of combinations, the new strategy eliminated the assignment of reads to indices not present in the sequencing mix (Table 3 and Supplementary table 5). This was the case regardless of whether the sequences were synthesised or not, suggesting that factory oligonucleotide cross-contamination is effectively eliminated by TruGrade synthesis. However, significant misassignment due to amplification or sequencing error remained a problem. As anticipated, we detected significant misassignment to indices where a frameshift had been introduced in the index. For example, even though the two index sequences have a high Hamming distance, frameshift between index i7_05 AACTCCGC and index i7_22 ACTCCGCA (Table 3) results in 0.6% of reads being misassigned. For subsequent experiments, we split the indexing oligonucleotides into groups

sharing compatible sequence combinations (Supplementary table 6), removing the potential for misassignment by frameshift error.

Our results show that the use of dual indices designed to maximise Hamming difference and minimise frameshift error reduces the risk of misassignment of samples in multiplexed MRD. However, it is apparent that even without frameshift error, up to 0.5% of reads are still assigned to non-existent index combinations for dual index reads (e.g. Table 3; i7_04/i5_06). This suggests similar errors must be present but hidden in the undetectable (mis)-assignment of reads to real indices using dual or single index strategies. This could clearly pose a concern for accurate MRD assessment.

MiSeq on-board demultiplexing software

On-board MiSeq Reporter software automatically demultiplexes MiSeq output from sequencing runs and converts binary base call (bcl) files to human readable text (fastq) files for each index on the sample sheet entered for the run. Clusters are assigned to a sample when the index sequence matches exactly but also permit assignment with a single mismatch per index read. We modified the process to assign only sample names to reads bearing a perfectly matched (i.e. 0 bp mismatch) index by demultiplexing raw sequencing output data using Illumina Consensus Assessment of Sequence and Variance (CASAVA) software version 1.8.2 (Illumina Inc., San Diego, CA) or bcl2fastq conversion software version 18.4 (Illumina Inc., San Diego, CA), allowing for no mismatches in the index sequence. The reanalysed data from run A7ELC is shown in Table 4 (and Supplementary table 7). As expected, there was a reduction in the misassigned reads (compare with Table 3), but also an unacceptable reduction in the assigned reads, with a correspondingly large increase in the undetermined bin. We noted that the quality statistics from the reads which remained misassigned were generally poor with %≥ Q30 lower for misassigned sequence reads. We therefore filtered the fastq files to remove any reads with % Q30 < 70 %, < 80 % and < 90 % (Supplementary table 8). As expected we lost a greater proportion of reads from the misassigned group than from the correctly assigned reads but the misassigned reads were still present and a large number of reads had been discarded – potentially impacting on sensitivity. We concluded that current demultiplexing strategies

247 are not stringent enough for very high sensitivity applications, and do not take account of the quality
248 of the index read.

249 ***Use of unique index combinations***

250 To further reduce the potential "hidden" misassignment, in the following run, A7FDO, we used only
251 unique combinations of group 1 indices (Supplementary table 6). We sequenced 8 samples, >1.5
252 million reads per sample, output data was demultiplexed using MiSeq on-board demultiplexer and
253 through CASAVA allowing for no mismatches in index reads, reducing overall misassignment from
254 0.2% to 0.05% (Supplementary tables 9 and 10). This shows that for MRD measurement where a
255 high degree of accuracy (minimising misassignment) is required only dual unique indices should be
256 used.

257 ***Demultiplexing based on quality of index reads – Error Aware Demultiplexer***

258 In view of the significant misassignment using the MiSeq on-board demultiplexer, the unacceptable
259 loss of potentially informative sequences when increasing stringency to allow no mismatches, poor
260 quality statistics of misassigned index reads, we developed our own demultiplexing pipeline. "Error-
261 Aware Demultiplexer" (EAD) utilises base call quality scores of index reads produced during
262 Illumina sequencing (open source available at: https://github.com/edm1/error-aware-demultiplexer).
263 The pipeline incorporates Phred scores to probabilistically match read indices to the sample identities
264 during demultiplexing. Index similarity is assessed with the same algorithm used in the Illumina pair-
265 end assembler PANDAseq [37]. The pipeline calculates the probability that the true index and the index
266 read represent the same underlying sequence. For example, if two bases match and the quality of
267 those bases are high then we have good evidence that they represent the same base. Probabilities are
268 calculated for each base and multiplied together to get the probability that the two reads represent the
269 same sequence. This results in up to an 80% reduction in misassignment compared to other
270 approaches and importantly, with minimal subsequent loss of sequence reads (Table 5).
271
272

Reducing run-to-run carryover

Template molecules may remain in the MiSeq fluidics system, even after a standard wash program, and can be washed onto the flow cell in subsequent runs. MiSeq instruments maintained according to standard Illumina recommendations typically have sample carryover rates of ≤ 0.1%. If the same indexing strategy is applied in a subsequent run, a 0.1% carryover rate could potentially cause errors in clinical interpretation of MRD. We added two non-human samples to the sequencing pool (A7ELC, Table 2) and then performed a sequencing run (A7FDO, Table 2) with the same indices but different samples 3 weeks later and after a further 2 different runs on the MiSeq (and therefore three standard washes). Carryover was detected at a rate of 0.002%. Performing the wash recipe recommended by Illumina for highly sensitive applications "Technical support: Reducing Run-to-Run Carryover on the MiSeq Using Dilute Sodium Hypochlorite Solution" (Illumina Inc., San Diego, CA), completely eradicated the carryover (data not shown). We observed that this wash recipe can sometimes cause lower cluster densities, presumably related to small amounts of sodium hypochlorite washed onto the flow cell.

Accurate quantification using a dPCR calculated reference "spike-in" – evidence for one in a million cells detection sensitivity

HTS-MRD is theoretically limited by the number of cells input. To accurately quantify sensitivity, we performed a "spike-in" experiment. While other groups have used plasmids or synthetic templates[18, 28], we used a pre-determined quantity of reference *IGH* DNA target derived from B-cell leukemia lines with unique *IGH* clonotypes extracted, prepared and accurately quantified using dPCR (data not shown). Dilutions ranging from $1x10^{-4}$ to $1x10^{-6}$ were created by adding known quantities (1, 10 and 100 copies) of SUPB15, TOM-1, and REH *IGH* sequences into one million cell equivalents of pooled donor lymphocyte DNA. We then amplified the DNA of all one million cells and sequenced the products using the HTS strategy described above. The sequencing run (ABG7Y, Table 2) was demultiplexed using EAD resulting in over 2.5 million reads per sample. We found using our workflow HTS-MRD achieved linear amplification, with $R^2 > 0.9998$, good reproducibility (Figure 4)

and the ability to detect the spike-in cell line *IGH* target down to one copy in a million normal cell equivalents.

Application to clinical MRD samples

To compare the effectiveness of our workflow with a standard HTS approach, we prepared libraries from five clinical samples previously scored for MRD by the gold standard ASO-PCR assay and classified into risk categories based on a 0.01% threshold. Each sample was indexed and sequenced together with 10^{-1} dilutions of its own diagnostic sample tagged with a different index. Demultiplexing and pre-processing errors in standard HTS workflows results in false positive risk classification due to the misassignment of diagnostic sample to patient MRD (Figure 5A). In the clinic this would lead to overtreatment of patients. The scenario was then repeated using the improved workflow with EAD. The samples were now all correctly classified (Figure 5B). This experiment demonstrates the potential clinical consequences of misclassification, and the power of an improved workflow to prevent this occurring even with a one in ten dilution of diagnostic sample present in the material.

Discussion

Several groups have demonstrated that HTS of the rearranged *IGH* gene is a potentially sensitive method for MRD detection in patients with ALL[15, 16, 18, 38]. However, adoption of new approaches in clinical laboratories requires careful evaluation of all potential sources of error and the development of strategies to ensure the highest accuracy[27].

Several important sources of potential error impact on workflow choice for HTS-MRD. Differential hybridisation kinetics of oligonucleotide primers can introduce significant biases that alter the composition of sequence libraries prepared by multiplex PCR[28]. However, for EuroMRD approved centres[10], the accepted clinical diagnostic assay for MRD is PCR based using consensus primers and it therefore seemed reasonable to adopt the same primer sets as the basis for the proposed HTS-MRD test. The Illumina MiSeq is subject to characteristic base-calling errors, but these are significantly lower than current competing systems. *Kennedy et al.*[39] describe a ligation and capture based assay

which overcome HTS errors, but the method is not suitable for introducing molecular indices into the specific locus for IG/TCR genes. Current techniques used for sensitive HTS-MRD are all amplicon-based approaches.

Multiplexing of indexed samples means that HTS-MRD could be an economical clinical assay. However, our analysis identified low, but clearly detectable and clinically relevant levels of sample misassignment using the standard MiSeq demultiplexing approach. We therefore developed a strategy that reduces misassignment to the absolute minimum while maintaining maximal sensitivity.

We adopted TruGrade-synthesised oligonucleotide primers designed with a high Hamming distance and screened to avoid frameshift error. *Quail et al.* reported contamination rates for purification by HPLC or PAGE purification of 0.56% and 0.34%, however with TruGrade this reduced to just 0.03%[31]. Also, for high sensitivity applications, unique combination dual-indexing, which identifies the sample origin of each sequence twice, independently, during demultiplexing, is superior to single-index multiplex sequencing[30].

Although misassigned reads are broadly of low quality, filtering based on read quality alone is an inefficient method for improving accuracy. A better method is to filter on index read quality, so we developed a custom demultiplexing pipeline (EAD) that uses a probabilistic approach to remove unreliable index reads while optimising retention of informative sequences. EAD out-performs standard demultiplexing software, including the on-board MiSeq demultiplexer, producing high quality data, with up to 80% reduction in read misassignment. Despite reduced tolerance for inaccurate index reads, EAD achieves up to 10% more allocated reads (Table 5) than other demultiplexing methods. EAD has potential application in any assay requiring highly accurate demultiplexing (e.g. pathogen detection and single cell applications). We also confirmed run-to-run carryover in the MiSeq fluidic pathway and demonstrated that it is essential to perform high-stringency post-run washes of the MiSeq to avoid the risk of contamination. Together, our workflow reduces misassignment to less than 0.05% with no loss of potentially usable data, resulting in high quality and accurately assigned data from multiplexed sequencing experiments.

We have highlighted the importance of eliminating all avoidable risk in MRD analysis, and appropriate quality control measures could provide some safeguards. *Seitz et al.*[40], for example, describe a novel method to prevent carry-over contaminations during similar two-step PCR protocols. It will also be good practice that no diagnostic sample should be sequenced together with its follow-up, that separate laboratory areas be used to prepare follow-up sample libraries, and that index combinations be alternated. Despite this, errors may still occur, chiefly associated with mis-identification of samples during multiplexing.

Firstly there is a significant risk that false-positive results may occur if two patients share exactly the same clonal sequence. Wu *et al.* evaluated this by HTS in post treatment samples and estimated the chance in B cells was 0.1% and 0.72% at 1 cell in 1 00 000 and 1 000 000 resolution respectively[16]. Next, the spike-in (cell lines, synthetic templates or plasmids) used for quantification (as in all current *IGH* HTS-MRD assays) will potentially all have the same clonal *IGH* sequence in each sample on the same run, risking inaccurate quantification with consequent potential for misclassification. Furthermore, simple human error in sample processing could result in misdiagnosis, or even where detected, the cost of a repeat run. We clearly demonstrate that a 10^{-1} diluted diagnostic sample can result in misclassification with the standard multiplexing approach, with the likelihood presumably diminishing with clone concentration. In the clinical setting, cost-effective diagnostic sequencing will depend on a high degree of multiplexing with optimal use of sequencing space on cheaper, higher capacity sequencers. In this scenario, cross-checking and excluding potentially contaminating highly expressed clones before sequencing would be inconvenient and reduce efficiency. It is therefore notable that our optimised approach with EAD reduces misassignment errors from all these scenarios and allows correct classification even with a 10% diagnostic sample present.

We have also introduced highly accurate and biologically relevant sequencing reference controls for HTS-MRD. Using dPCR, we absolutely quantified the number of molecules of biologically relevant *IGH* controls present in each preparation. This is more accurate than conventional RQ-PCR or estimations based on molecular weight and concentration. The entire workflow was then applied to demonstrate sensitivity of at least one in a million cell equivalents, improving sensitivity by at least

ten fold compared to the current ASO-RQ MRD approach. A major advantage of such sensitivity, combined with confident sample assignment in multiplexing would be to allow application to peripheral blood rather than BM, with obvious advantages for patients (especially children), as well as potentially giving a more representative picture of MRD[41-46].

The improvements described represent a step towards the rigorous validation required to produce a robust clinical HTS-MRD assay. Large prospective head-to-head comparison with current methods are needed to prove if the increased sensitivity and broader view of IG repertoire that HTS can achieve can replace the current burdensome and expensive techniques. We also propose EAD as a flexible method applicable beyond MRD detection to any multiplexing approaches where high accuracy of assignment is paramount.

Acknowledgments

This work was supported by Great Ormond Street Hospital Children's Charity. We wish to give special thanks to the Smurfit family for the generous support of this MRD research. The work described forms part of the EuroClonality-NGS consortium to standardise workflows. Finally we would like to acknowledge Above and Beyond, Charitable Trustees of University Hospitals Bristol for their funding which supported the original design of the MRD analysis pipeline.

Authorship contribution: J.B., M.H. and N.G. designed the research. J.B., G.W., and T.B. performed experiments; E.M. designed bioinformatics pipelines with help of J.B., H.W., J.H., J.M. J.B. and M.H.; J.B. and M.H. analysed results and J.B. made the figures; J.B. and M.H. wrote the paper. E.M., T.B., G.W., H.W., J.H., N.G. and J.M. contributed to the writing of the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Dr Mike Hubank, Genetics and Genomic Medicine Programme, UCL Institute of Child Health, London. E-mail: m.hubank@ucl.ac.uk

References

[1] Campana D: Role of minimal residual disease monitoring in adult and pediatric acute lymphoblastic leukemia. Hematology/oncology clinics of North America 2009, 23:1083-98, vii.

[2] Yeoh AE, Ariffin H, Chai EL, Kwok CS, Chan YH, Ponnudurai K, Campana D, Tan PL, Chan MY, Kham SK, Chong LA, Tan AM, Lin HP, Quah TC: Minimal residual disease-guided treatment deintensification for children with acute lymphoblastic leukemia: results from the Malaysia-Singapore acute lymphoblastic leukemia 2003 study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012, 30:2384-92.

[3] Campana D: Minimal residual disease monitoring in childhood acute lymphoblastic leukemia. Current opinion in hematology 2012, 19:313-8.

[4] Vora A, Goulden N, Wade R, Mitchell C, Hancock J, Hough R, Rowntree C, Richards S: Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. The lancet oncology 2013, 14:199-209.

[5] Conter V, Bartram CR, Valsecchi MG, Schrauder A, Panzer-Grumayer R, Moricke A, Arico M, Zimmermann M, Mann G, De Rossi G, Stanulla M, Locatelli F, Basso G, Niggli F, Barisone E, Henze G, Ludwig WD, Haas OA, Cazzaniga G, Koehler R, Silvestri D, Bradtke J, Parasole R, Beier R, van Dongen JJ, Biondi A, Schrappe M: Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. Blood 2010, 115:3206-14.

[6] Yamaji K, Okamoto T, Yokota S, Watanabe A, Horikoshi Y, Asami K, Kikuta A, Hyakuna N, Saikawa Y, Ueyama J, Watanabe T, Okada M, Taga T, Kanegane H, Kogawa K, Chin M, Iwai A, Matsushita T, Shimomura Y, Hori T, Tsurusawa M: Minimal residual disease-based augmented therapy in childhood acute lymphoblastic leukemia: a report from the Japanese Childhood Cancer and Leukemia Study Group. Pediatric blood & cancer 2010, 55:1287-95.

427 [7] Goulden NJ, Knechtli CJ, Garland RJ, Langlands K, Hancock JP, Potter MN, Steward CG, Oakhill A:
428 Minimal residual disease analysis for the prediction of relapse in children with standard-risk acute
429 lymphoblastic leukaemia. British journal of haematology 1998, 100:235-44.
430 [8] van Dongen JJ, Seriu T, Panzer-Grumayer ER, Biondi A, Pongers-Willemse MJ, Corral L, Stolz F,
431 Schrappe M, Masera G, Kamps WA, Gadner H, van Wering ER, Ludwig WD, Basso G, de Bruijn MA,
432 Cazzaniga G, Hettinger K, van der Does-van den Berg A, Hop WC, Riehm H, Bartram CR: Prognostic
433 value of minimal residual disease in acute lymphoblastic leukaemia in childhood. Lancet 1998,
434 352:1731-8.
435 [9] Gaipa G, Basso G, Biondi A, Campana D: Detection of minimal residual disease in pediatric acute
436 lymphoblastic leukemia. Cytometry Part B, Clinical cytometry 2013.
437 [10] van der Velden VH, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grumayer ER, Flohr T,
438 Sutton R, Cave H, Madsen HO, Cayuela JM, Trka J, Eckert C, Foroni L, Zur Stadt U, Beldjord K, Raff T,
439 van der Schoot CE, van Dongen JJ: Analysis of minimal residual disease by Ig/TCR gene
440 rearrangements: guidelines for interpretation of real-time quantitative PCR data. Leukemia 2007,
441 21:604-11.
442 [11] Kalina T, Flores-Montero J, van der Velden VH, Martin-Ayuso M, Bottcher S, Ritgen M, Almeida J,
443 Lhermitte L, Asnafi V, Mendonca A, de Tute R, Cullen M, Sedek L, Vidriales MB, Perez JJ, te Marvelde
444 JG, Mejstrikova E, Hrusak O, Szczepanski T, van Dongen JJ, Orfao A: EuroFlow standardization of flow
445 cytometer instrument settings and immunophenotyping protocols. Leukemia 2012, 26:1986-2010.
446 [12] Szczepanski T, Willemse MJ, Brinkhof B, van Wering ER, van der Burg M, van Dongen JJ:
447 Comparative analysis of Ig and TCR gene rearrangements at diagnosis and at relapse of childhood
448 precursor-B-ALL provides improved strategies for selection of stable PCR targets for monitoring of
449 minimal residual disease. Blood 2002, 99:2315-23.
450 [13] Kohlmann A, Grossmann V, Haferlach T: Integration of next-generation sequencing into clinical
451 practice: are we there yet? Seminars in oncology 2012, 39:26-36.

452 [14] Logan AC, Zhang B, Narasimhan B, Carlton V, Zheng J, Moorhead M, Krampf MR, Jones CD,
453 Waqar AN, Faham M, Zehnder JL, Miklos DB: Minimal residual disease quantification using
454 consensus primers and high-throughput IGH sequencing predicts post-transplant relapse in chronic
455 lymphocytic leukemia. Leukemia 2013, 27:1659-65.
456 [15] Ladetto M, Bruggemann M, Monitillo L, Ferrero S, Pepin F, Drandi D, Barbero D, Palumbo A,
457 Passera R, Boccadoro M, Ritgen M, Gokbuget N, Zheng J, Carlton V, Trautmann H, Faham M, Pott C:
458 Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in
459 B-cell disorders. Leukemia 2014, 28:1299-307.
460 [16] Wu D, Emerson RO, Sherwood A, Loh ML, Angiolillo A, Howie B, Vogt J, Rieder M, Kirsch I,
461 Carlson C, Williamson D, Wood BL, Robins H: Detection of Minimal Residual Disease in B
462 Lymphoblastic Leukemia by High-Throughput Sequencing of IGH. Clinical cancer research : an official
463 journal of the American Association for Cancer Research 2014, 20:4540-8.
464 [17] Logan AC, Vashi N, Faham M, Carlton V, Kong K, Buno I, Zheng J, Moorhead M, Klinger M, Zhang
465 B, Waqar A, Zehnder JL, Miklos DB: Immunoglobulin and T cell receptor gene high-throughput
466 sequencing quantifies minimal residual disease in acute lymphoblastic leukemia and predicts post-
467 transplantation relapse and survival. Biology of blood and marrow transplantation : journal of the
468 American Society for Blood and Marrow Transplantation 2014, 20:1307-13.
469 [18] Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, Pui CH, Campana D:
470 Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia.
471 Blood 2012, 120:5173-80.
472 [19] Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buno I, Armstrong R, Fire AZ,
473 Weinberg KI, Mindrinos M, Zehnder JL, Boyd SD, Xiao W, Davis RW, Miklos DB: High-throughput VDJ
474 sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and
475 immune reconstitution assessment. Proceedings of the National Academy of Sciences of the United
476 States of America 2011, 108:21194-9.

[20] Martinez-Lopez J, Lahuerta JJ, Pepin F, Gonzalez M, Barrio S, Ayala R, Puig N, Montalban MA, Paiva B, Weng L, Jimenez C, Sopena M, Moorhead M, Cedena T, Rapado I, Mateos MV, Rosinol L, Oriol A, Blanchard MJ, Martinez R, Blade J, San Miguel J, Faham M, Garcia-Sanz R: Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma. Blood 2014, 123:3073-9.

[21] Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, Greisman HA, Sabath DE, Wood BL, Robins H: High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Science translational medicine 2012, 4:134ra63.

[22] Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Muller CR, Pratt V, Wallace A: A standardized framework for the validation and verification of clinical molecular genetic tests. European journal of human genetics : EJHG 2010, 18:1276-88.

[23] Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ: Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. Journal of pathology informatics 2012, 3:40.

[24] Williams ES, Hegde M: Implementing genomic medicine in pathology. Advances in anatomic pathology 2013, 20:238-44.

[25] Xuan J, Yu Y, Qing T, Guo L, Shi L: Next-generation sequencing in the clinic: promises and challenges. Cancer letters 2013, 340:284-95.

[26] Nekrutenko A, Taylor J: Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nature reviews Genetics 2012, 13:667-72.

[27] van Dongen JJ, van der Velden VH, Bruggemann M, Orfao A: Minimal residual disease (MRD) diagnostics in acute lymphoblastic leukemia (ALL): need for sensitive, fast and standardized technologies. Blood 2015.

[28] Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, Wu D, Wood BL, Rieder MJ, Robins H:

502 Using synthetic templates to design an unbiased multiplex PCR assay. Nature communications 2013,
503 4:2680.
504 [29] Collins FS, Hamburg MA: First FDA authorization for next-generation sequencer. The New
505 England journal of medicine 2013, 369:2369-71.
506 [30] Kircher M, Sawyer S, Meyer M: Double indexing overcomes inaccuracies in multiplex sequencing
507 on the Illumina platform. Nucleic acids research 2012, 40:e3.
508 [31] Quail MA, Smith M, Jackson D, Leonard S, Skelly T, Swerdlow HP, Gu Y, Ellis P: SASI-Seq: sample
509 assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. BMC genomics
510 2014, 15:110.
511 [32] Verhagen OJ, Wijkhuijs AJ, van der Sluijs-Gelling AJ, Szczepanski T, van der Linden-Schrever BE,
512 Pongers-Willemse MJ, van Wering ER, van Dongen JJ, van der Schoot CE: Suitable DNA isolation
513 method for the detection of minimal residual disease by PCR techniques. Leukemia 1999, 13:1298-9.
514 [33] Kozarewa I, Turner DJ: 96-plex molecular barcoding for the Illumina Genome Analyzer. Methods
515 Mol Biol 2011, 733:279-98.
516 [34] Bhat S, Herrmann J, Armishaw P, Corbisier P, Emslie KR: Single molecule detection in nanofluidic
517 digital array enables accurate measurement of DNA copy number. Analytical and bioanalytical
518 chemistry 2009, 394:457-67.
519 [35] Mir K, Neuhaus K, Bossert M, Schober S: Short barcodes for next generation sequencing. PloS
520 one 2013, 8:e82933.
521 [36] Bystrykh LV: Generalized DNA barcode design based on Hamming codes. PloS one 2012,
522 7:e36852.
523 [37] Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD: PANDAseq: paired-end
524 assembler for illumina sequences. BMC bioinformatics 2012, 13:31.
525 [38] Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A, Grardel N, Roumier C, Preudhomme
526 C, Figeac M: Fast multiclonal clusterization of V(D)J recombinations from high-throughput
527 sequencing. BMC genomics 2014, 15:409.

528 [39] Kennedy SR, Schmitt MW, Fox EJ: Detecting ultralow-frequency mutations by Duplex
529 Sequencing. 2014, 9:2586-606.
530 [40] Seitz V, Schaper S, Droge A, Lenze D, Hummel M, Hennig S: A new method to prevent carry-over
531 contaminations in two-step PCR NGS library preparations. Nucleic acids research 2015.
532 [41] Brisco MJ, Sykes PJ, Hughes E, Dolman G, Neoh SH, Peng LM, Toogood I, Morley AA: Monitoring
533 minimal residual disease in peripheral blood in B-lineage acute lymphoblastic leukaemia. British
534 journal of haematology 1997, 99:314-9.
535 [42] Brisco MJ, Sykes PJ, Hughes E, Story CJ, Rice MS, Schwarer AP, Morley AA: Molecular relapse can
536 be detected in blood in a sensitive and timely fashion in B-lineage acute lymphoblastic leukemia.
537 Leukemia 2001, 15:1801-2.
538 [43] Coustan-Smith E, Sancho J, Hancock ML, Razzouk BI, Ribeiro RC, Rivera GK, Rubnitz JE, Sandlund
539 JT, Pui CH, Campana D: Use of peripheral blood instead of bone marrow to monitor residual disease
540 in children with acute lymphoblastic leukemia. Blood 2002, 100:2399-402.
541 [44] Helgestad J, Rosthoj S, Johansen P, Varming K, Ostergaard E: Bone marrow aspiration technique
542 may have an impact on therapy stratification in children with acute lymphoblastic leukaemia.
543 Pediatric blood & cancer 2011, 57:224-6.
544 [45] Martin H, Atta J, Bruecher J, Elsner S, Schardt C, Stadler M, von Melchner H, Hoelzer D: In
545 patients with BCR-ABL-positive ALL in CR peripheral blood contains less residual disease than bone
546 marrow: implications for autologous BMT. Annals of hematology 1994, 68:85-7.
547 [46] van der Velden VH, Jacobs DC, Wijkhuijs AJ, Comans-Bitter WM, Willemse MJ, Hahlen K, Kamps
548 WA, van Wering ER, van Dongen JJ: Minimal residual disease levels in bone marrow and peripheral
549 blood are comparable in children with T cell acute lymphoblastic leukemia (ALL), but not in
550 precursor-B-ALL. Leukemia 2002, 16:1432-6.

551

552

Tables

Table 1. Oligonucleotide designs for dual-indexing. These are used in the second stage of a nested PCR. First stage primers have adaptors† at each end of amplicon which are complementary to the indexing oligonucleotide.

Index	Oligonucleotide Sequence	Index in oligonucleotide	Index read sequence
i7_01	5'-CAAGCAGAAGACGGCATACGAGATTCTAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TCTAGCTA	TAGCTAGA
i7_02	5'-CAAGCAGAAGACGGCATACGAGATCTAGCTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	CTAGCTAT	ATAGCTAG
i7_03	5'-CAAGCAGAAGACGGCATACGAGATAGGTTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	AGGTTGGC	GCCAACCT
i7_04	5'-CAAGCAGAAGACGGCATACGAGATGACCAACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	GACCAACG	CGTTGGTC
i7_05	5'-CAAGCAGAAGACGGCATACGAGATGCGGAGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	GCGGAGTT	AACTCCGC
i7_06	5'-CAAGCAGAAGACGGCATACGAGATGTGCCATAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	GTGCCATA	TATGGCAC
i7_07	5'-CAAGCAGAAGACGGCATACGAGATTAATGTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TAATGTCC	GGACATTA
i7_08	5'-CAAGCAGAAGACGGCATACGAGATCGAAGGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	CGAAGGAC	GTCCTTCG
i7_09	5'-CAAGCAGAAGACGGCATACGAGATAATGTCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	AATGTCCT	AGGACATT
i7_10	5'-CAAGCAGAAGACGGCATACGAGATAGAACATTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	AGAACATT	AATGTTCT
i7_11	5'-CAAGCAGAAGACGGCATACGAGATTGTCAGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TGTCAGTC	GACTGACA
i7_12	5'-CAAGCAGAAGACGGCATACGAGATCACCGCTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	CACCGCTT	AAGCGGTG
i7_13	5'-CAAGCAGAAGACGGCATACGAGATCAGACGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	CAGACGCA	TGCGTCTG
i7_14	5'-CAAGCAGAAGACGGCATACGAGATGCTACTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	GCTACTAG	CTAGTAGC
i7_15	5'-CAAGCAGAAGACGGCATACGAGATGTCAGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	GTCAGTCT	AGACTGAC
i7_16	5'-CAAGCAGAAGACGGCATACGAGATTTCACCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TTCACCGC	GCGGTGAA
i7_17	5'-CAAGCAGAAGACGGCATACGAGATGGTCTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	GGTCTAAT	ATTAGACC
i7_18	5'-CAAGCAGAAGACGGCATACGAGATACCTGGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	ACCTGGAT	ATCCAGGT
i7_19	5'-CAAGCAGAAGACGGCATACGAGATAGCGACAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	AGCGACAG	CTGTCGCT
i7_20	5'-CAAGCAGAAGACGGCATACGAGATATAGGCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	ATAGGCTC	GAGCCTAT
i7_21	5'-CAAGCAGAAGACGGCATACGAGATTAGAACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TAGAACAT	ATGTTCTA
i7_22	5'-CAAGCAGAAGACGGCATACGAGATTGCGGAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TGCGGAGT	ACTCCGCA
i7_23	5'-CAAGCAGAAGACGGCATACGAGATTTGCGGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TTGCGGAG	CTCCGCAA
i7_24	5'-CAAGCAGAAGACGGCATACGAGATTTAGAACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T-3'	TTAGAACA	TGTTCTAA
i5_01	5'-AATGATACGGCGACCACCGAGATCTACACCACTTGAGACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	CACTTGAG	CACTTGAG
i5_02	5'-AATGATACGGCGACCACCGAGATCTACACGTTACCGAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	GTTACCGA	GTTACCGA
i5_03	5'-AATGATACGGCGACCACCGAGATCTACACTGACGACTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	TGACGACT	TGACGACT
i5_04	5'-AATGATACGGCGACCACCGAGATCTACACACGGATTCACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	ACGGATTC	ACGGATTC
i5_05	5'-AATGATACGGCGACCACCGAGATCTACACCCATAGGAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	CCATAGGA	CCATAGGA
i5_06	5'-AATGATACGGCGACCACCGAGATCTACACTGGAAGGCACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	TGGAAGGC	TGGAAGGC
i5_07	5'-AATGATACGGCGACCACCGAGATCTACACGCATCATGACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	GCATCATG	GCATCATG
i5_08	5'-AATGATACGGCGACCACCGAGATCTACACAGCGGTGAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	AGCGGTGA	AGCGGTGA
i5_09	5'-AATGATACGGCGACCACCGAGATCTACACAGTTACCGACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	AGTTACCG	AGTTACCG
i5_10	5'-AATGATACGGCGACCACCGAGATCTACACCATGCATAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	CATGCATA	CATGCATA
i5_11	5'-AATGATACGGCGACCACCGAGATCTACACACATGCATACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	ACATGCAT	ACATGCAT
i5_12	5'-AATGATACGGCGACCACCGAGATCTACACACCATAGGACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	ACCATAGG	ACCATAGG
i5_13	5'-AATGATACGGCGACCACCGAGATCTACACTCCAGGTAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	TCCAGGTA	TCCAGGTA
i5_14	5'-AATGATACGGCGACCACCGAGATCTACACCTTAATTGACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	CTTAATTG	CTTAATTG
i5_15	5'-AATGATACGGCGACCACCGAGATCTACACCGGATTCAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	CGGATTCA	CGGATTCA
i5_16	5'-AATGATACGGCGACCACCGAGATCTACACTTAGACCAACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'	TTAGACCA	TTAGACCA

†First stage primers including partial adaptor sequences for - Forward (i7) primers:
5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGGGTGCGACAGGCCCCTGGACAA-3'
5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGGATCCGTCAGCCCCCAGGGAAGG-3'
5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGTCCGCCAGGCTCCAGGGAA-3'
5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGGATCCGCCAGCCCCCAGGGAAGG-3'
5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGTGCGCCAGATGCCCGGGAAAGG-3'

5’-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGGATCAGGCAGTCCCCATCGAGAG-3’
5’-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTGGGTGCGACAGGCCCCTGGACAA-3’
Reverse primer (i5):
5’-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTACCTGAGGAGACGGTGACC-3’
*Indicates addition of phosphorothioated DNA bases

568 Table 2. Sequencing runs performed on the Illumina MiSeq

MiSeq Run	Unique run name	Number of samples	Indexing strategy	MiSeq kit version	Read length	Total reads - million	Cluster density (K/mm^2)	Reads passing filter- million	Phred quality (%>=Q30)
1	A478Y	36	Single, Kozarewa	2	300	22.4	1211	17.38	71.0
2	A5U9G	34	Single, Kozarewa	2	300	22.12	1094	18.69	74.4
3	A6PKD	20	Single, Kozarewa	2	300	20.91	1006	17.55	78.6
4	A6FMV	23	Single, Kozarewa	3	167	12.11	465	11.70	95.9
5	A7BK7	28	Single, Kozarewa	2	208	22.35	1063	18.59	79.2
6	A7ELC	31	Dual, TruGrade	3	151	32.83	1267	22.81	80.3
7	A7FDO	8	Dual, TruGrade	3	151	23.98	865	18.77	82.1
8	A72MP	8	Dual, TruGrade	2	300	23.01	1067	20.17	70.1
9	A7B79	21	Dual, TruGrade	2	300	24.45	1175	19.66	72.5
10	A8FPT	12	Dual, TruGrade	2	300	21.66	1020	19.53	78.6
11	ABALR	12	Dual, TruGrade	2	300	23.44	1120	18.33	72.6
12	ABAJL	37	Dual, TruGrade	2	300	24.19	1170	21.77	69.3
13	ABG7Y	6	Dual, TruGrade	3	151	24.29	892	22.48	85

569

570

Table 3. Assignment of reads from dual-indexed run A7ELC using the MiSeq on-board demultiplexer. Index combinations corresponding to actual samples are highlighted in bold. Misassigned reads (non-bold print) are shown for all possible dual-index combinations where at least one index is assigned (complete information in Supplementary table 5). Index 1 has the prefix i5 and index 2 has the prefix i7. An example of frameshift causing misassignment occurs between indices i7_05 and i7_22 (underlined text) resulting in misassignment of 0.6% of reads.

		i7 indices															
		i7_01	*i7_02*	*i7_03*	*i7_04*	*i7_05*	*i7_06*	*i7_07*	*i7_08*	*i7_09*	*i7_14*	*i7_17*	*i7_19*	*i7_20*	*i7_22*	*i7_23*	*i7_24*
i5 indices	*i5_01*	**532594**	**553987**	**606447**	2105	***715211***	**586926**	3581	**500618**	0	2	5	2	4	***2642***	140	0
	i5_02	**483605**	**420978**	3407	**563981**	*2130*	**694559**	**457959**	1129	153	4	7	3	0	***2***	0	3
	i5_03	4306	**413460**	**853442**	**912389**	***542539***	2107	**785862**	**433723**	234	0	6	0	16	***1851***	103	0
	i5_04	**869143**	2301	**820461**	**485673**	***708010***	**548037**	3374	**501739**	2	0	0	0	10	***2470***	162	0
	i5_05	5541	**398969**	4490	**794442**	***608924***	3922	**631108**	1442	170	0	3	3	0	***1949***	97	2
	i5_06	**1410284**	3156	**807859**	2536	*2938*	**731447**	**613956**	1927	189	0	0	0	14	***5***	0	0
	i5_09	163	131	0	186	*0*	223	123	0	0	0	0	0	0	***0***	0	0
	i5_10	0	2	0	0	*0*	0	0	2	0	0	0	0	0	***0***	0	0
	i5_12	1	68	0	145	*130*	0	127	1	0	0	0	0	0	***0***	0	0
	i5_14	0	0	1	0	*0*	0	0	0	0	0	0	0	0	***0***	0	0
	i5_15	803	1	835	443	*718*	556	6	543	0	0	0	0	0	***5***	0	0
	i5_16	4	6	8	2	*7*	2	9	2	0	0	0	0	0	***0***	0	0
	Undetermined reads 3E+06																

Table 4. Assignment of reads from dual-indexed run A7ELC demultiplexed allowing for no mismatches in index sequences. Index combinations for samples included in the run are shown in bold print. Misassigned reads (non-bold print) are shown for all possible dual-index combinations where at least one index is assigned (complete information in Supplementary table 7). Index 1 has the prefix i5 and index 2 has the prefix i7.

		i7 indices									
		i7_01	*i7_02*	*i7_03*	*i7_04*	*i7_05*	*i7_06*	*i7_07*	*i7_08*	*i7_22*	*i7_23*
i5 indices	*i5_01*	**484180**	**505456**	**549076**	402	**654781**	**535588**	661	**449768**	1689	12
	i5_02	**436266**	**387331**	741	**511368**	440	**629466**	**415014**	241	0	0
	i5_03	1138	**376242**	**790058**	**893669**	**489608**	472	**714305**	**394937**	1177	10
	i5_04	**809651**	454	**738566**	**439508**	**690965**	**490128**	683	**450862**	1715	2
	i5_05	1218	**358552**	970	**722159**	**555629**	637	**569741**	250	1239	10
	i5_06	**1291695**	535	**733096**	525	614	**660168**	**552237**	360	2	0
	i5_09	0	0	0	0	0	1	0	0	0	0
	i5_12	0	1	0	1	1	0	1	0	0	0
	i5_15	561	0	583	298	522	375	0	365	4	0
	i5_16	0	0	0	0	0	0	0	0	0	0
	Undetermined reads 5E+06										

Table 5. Reduction of misassignment due to index read sequence error using the Error Aware Demultiplexing strategy. Numbers of reads assigned, with percentage change in reads assigned compared to standard MiSeq on-board demultiplexing, allowing up to 1 base pair mismatch (in brackets); Reads assigned using standard methods allowing for 0 base pair mismatches in index reads are on upper rows (0 bp) of each index combination and Error Aware Demultiplexer read assignments are on the lower rows (EAD) of each index combination. The bold text indicate sample index combinations that were present in sequenced samples, the non-bold are index combinations not included in the sequencing run. Data from sequencing run A7FDO. Index 1 has the prefix i5 and index 2 has the prefix i7.

			i7 indices							
			i7_01	*i7_03*	*i7_04*	*i7_05*	*i7_06*	*i7_07*	*i7_08*	*i7_10*
i5 indices	*i5_01*	0 bp	**1633722 (-8.6)**	44 (-76.2)	76 (-43.3)	132 (-68.3)	109 (-72.5)	118 (-66.9)	86 (-74.9)	108 (-73.3)
		EAD	**1761443 (-1.4)**	89 (-51.9)	63 (-53.0)	102 (-75.5)	140 (-64.6)	110 (-69.2)	93 (-72.9)	85 (-79.0)
	i5_02	0 bp	154 (-50.2)	**1714456 (-8.8)**	121 (-46.5)	152 (-72.0)	169 (-50.4)	165 (-76.9)	129 (-81.8)	156 (-54.5)
		EAD	124 (-59.9)	**1921136 (-2.2)**	87 (-61.5)	127 (-76.6)	177 (-48.1)	174 (-75.6)	129 (-81.8)	181 (-47.2)
	i5_03	0 bp	50 (-60.0)	79 (-42.8)	**1506573 (-8.5)**	92 (-85.8)	59 (-67.6)	68 (-89.4)	61 (-69.3)	65 (-85.2)
		EAD	37 (-70.4)	73 (-47.1)	**1620923 (-1.5)**	63 (-90.3)	36 (-80.2)	55 (-91.4)	62 (-68.8)	55 (-87.5)
	i5_04	0 bp	86 (-54.0)	84 (-45.5)	60 (-37.5)	**1667394 (-9.0)**	75 (-58.3)	61 (-72.9)	80 (-80.5)	90 (-88.9)
		EAD	71 (-62.0)	76 (-50.6)	52 (-45.8)	**1800250 (-1.7)**	64 (-64.4)	62 (-72.4)	73 (-82.2)	101 (-87.6)
	i5_05	0 bp	185 (-82.8)	171 (-60.7)	150 (-66.7)	254 (-87.3)	**1748758 (-9.0)**	158 (-82.3)	118 (-78.9)	215 (-88.6)
		EAD	194 (-81.9)	189 (-56.6)	115 (-74.4)	168 (-91.6)	**1909784 (-0.6)**	169 (-81.1)	110 (-80.3)	180 (-90.5)
	i5_06	0 bp	185 (-64.4)	193 (-73.9)	154 (-79.8)	185 (-80)	170 (-57.5)	**1924357 (-8.2)**	160 (-60.7)	161 (-55.6)
		EAD	159 (-69.4)	160 (-78.3)	93 (-87.8)	173 (-81.3)	179 (-55.3)	**2083579 (-0.6)**	127 (-68.8)	164 (-54.8)
	i5_07	0 bp	152 (-70.9)	222 (-79.7)	130 (-60.2)	198 (-76.1)	167 (-70.1)	137 (-67.0)	**1586001 (-9.1)**	172 (-69.4)
		EAD	138 (-73.6)	206 (-81.1)	147 (-55.0)	174 (-79.0)	157 (-71.9)	134 (-67.7)	**1765359 (-1.1)**	152 (-73.0)
	i5_08	0 bp	44 (-86.3)	57 (-87.8)	40 (-79.9)	98 (-93.4)	60 (-92.2)	55 (-85.7)	31 (-88.0)	**1781666 (-8.8)**
		EAD	32 (-90.0)	58 (-87.6)	38 (-80.9)	74 (-95.0)	46 (-94.1)	47 (-87.8)	30 (-88.4)	**1979346 (-1.3)**

593 **Figure Legends**

594 **Figure 1. Nested PCR and final library product for *IGH* VDJ amplicon sequencing on the**
595 **Illumina MiSeq.** (A) Family variable heavy chain segment (VH) and joining heavy chain segment
596 (JH) primers with complementary partial index sequences are used to amplify the VDJ junction of the
597 rearranged *IGH* gene, containing the hypervariable region (purple). (B) In the second stage, index
598 sequences (blue) and Illumina P5 and P7 platform adaptors sequences are added. The final amplicon
599 construct with sequencing strategy is shown in (C). One or two index reads were used depending on
600 indexing strategy, with single-end sequencing read from the JH (P5) end of the amplicon.

601 **Figure 2. Misassignment of sequences from sequencing run A7BK7.** Indexed reads corresponding
602 to samples are shown in red. Reads assigned to indices not included in the sequencing run are shown
603 in blue. Eight bp indices used as per *Kowenza et al*.[33]

604 **Figure 3. Comparison of quality statistics between reads assigned to real samples ("true") and**
605 **reads "misassigned" to indices not included in the run (examples using data from run A7ELC).**
606 (A-C) Base quality of sequencing reads shown by position (A) mean quality score distribution for all
607 sequences (B) and index read quality (C) from a representative "true" indexed sample. (D-F) Base
608 quality of sequencing reads shown by position (D) mean quality score distribution for all sequences
609 (E) and index read quality (F) from a representative "misassigned" sample. Pooled libraries were
610 sequenced (150 bases, single-end) on an Illumina MiSeq.

611 **Figure 4. Accurate quantification of MRD down to 1 in 1 million cells using a spike in control**
612 **quantified by digital PCR.** Three separate serial dilutions of a cell lines; (A) SUPB15, (B) TOM-1
613 and (C) REH, spiked into to 1 million cells equivalent of pooled normal lymphocyte DNA. Samples
614 sequenced on Illumina MiSeq (> 2 million reads per sample), indexed using HTS strategy described
615 above and demultiplexed using Error Aware Demultiplexer.

616 **Figure 5. MRD analysis performed in 5 patients at the end of induction chemotherapy for**
617 **childhood B-ALL (day 28 MRD).** The current gold standard RQ-PCR technique is compared to a
618 standard HTS work flow (A) and an optimised workflow using Error Aware Demultiplexer (B). The

sequencing run included a diluted (10^{-1}) diagnostic sample for each patient to simulate an increased probability of misassignment. MRD samples were also "spiked" with cell line DNA for quantification purposes. Errors introduced during standard multiplexed sequencing result in mis-diagnosis in patients N001 and N005 due to misassignment (A). These incorrect calls due to misassignment are removed (B) using the workflow presented.

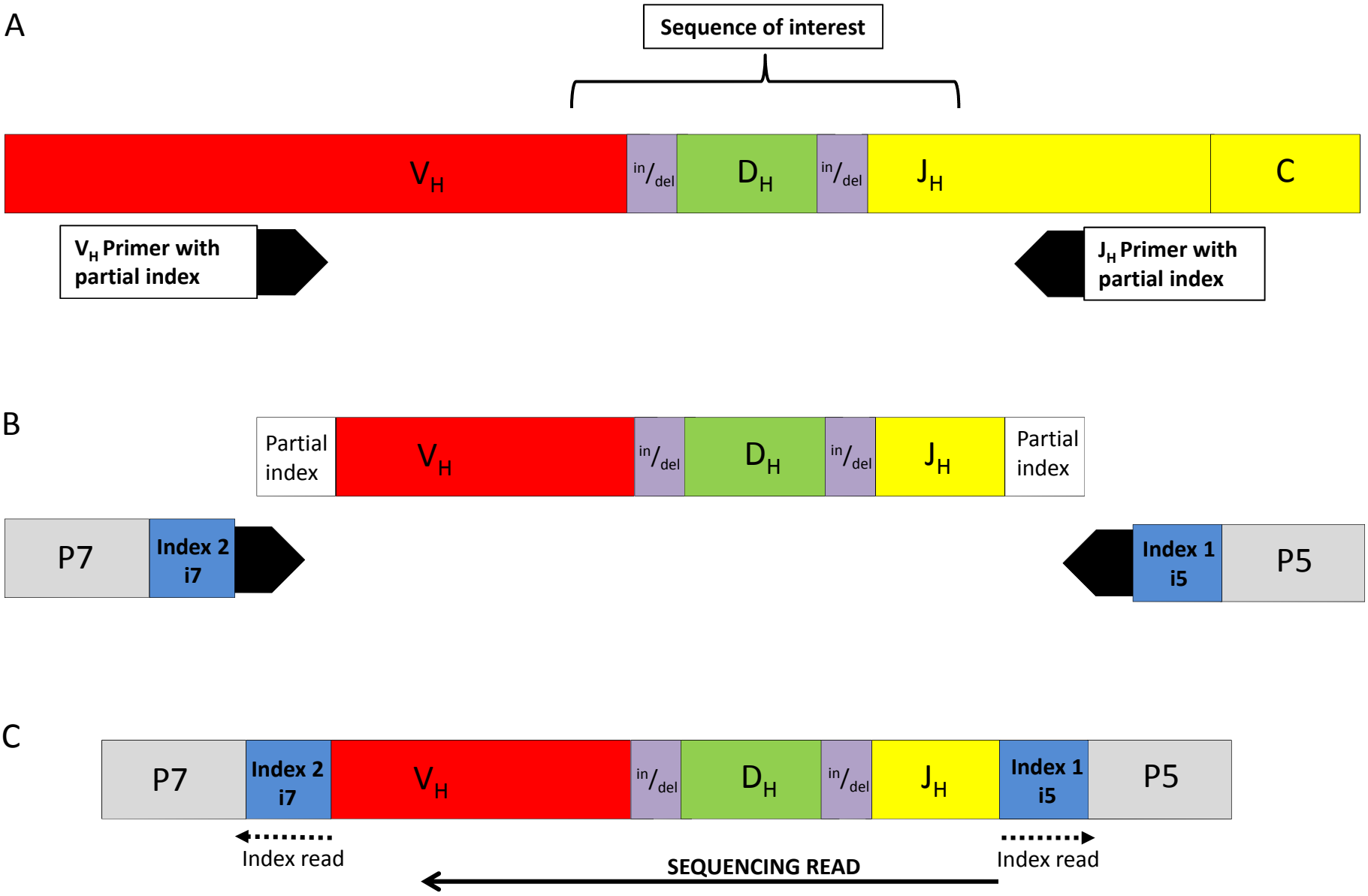
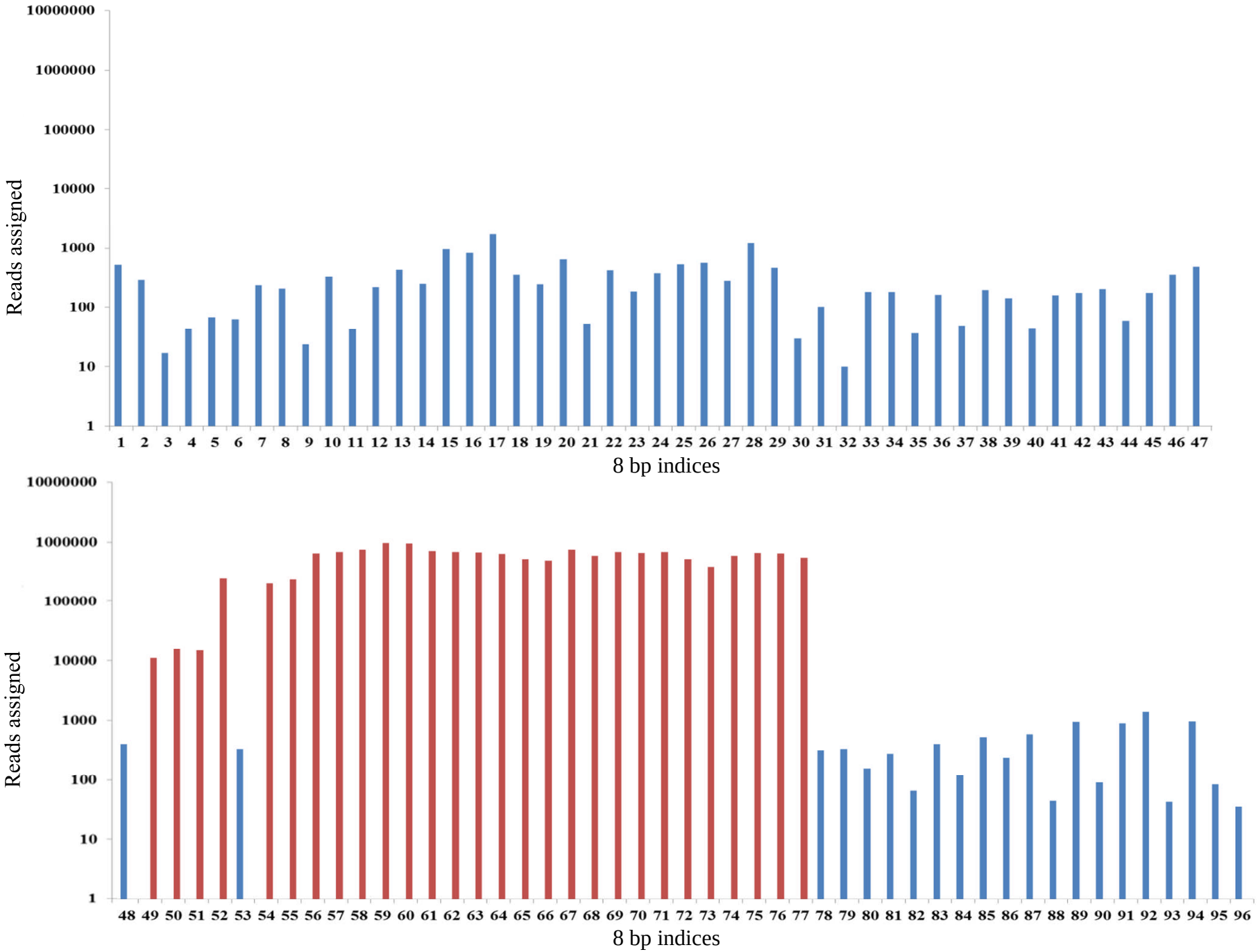

A
Sequence of interest
V_H
in/del
D_H
in/del
J_H
C
V_H Primer with partial index
J_H Primer with partial index
B
Partial index
V_H
in/del
D_H
in/del
J_H
Partial index
P7
Index 2 i7
Index 1 i5
P5
C
P7
Index 2 i7
V_H
in/del
D_H
in/del
J_H
Index 1 i5
P5
Index read
SEQUENCING READ
Index read

Figure 1.


Reads assigned
10000000
1000000
100000
10000
1000
100
10
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
8 bp indices
Reads assigned
10000000
1000000
100000
10000
1000
100
10
1
48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96
8 bp indices

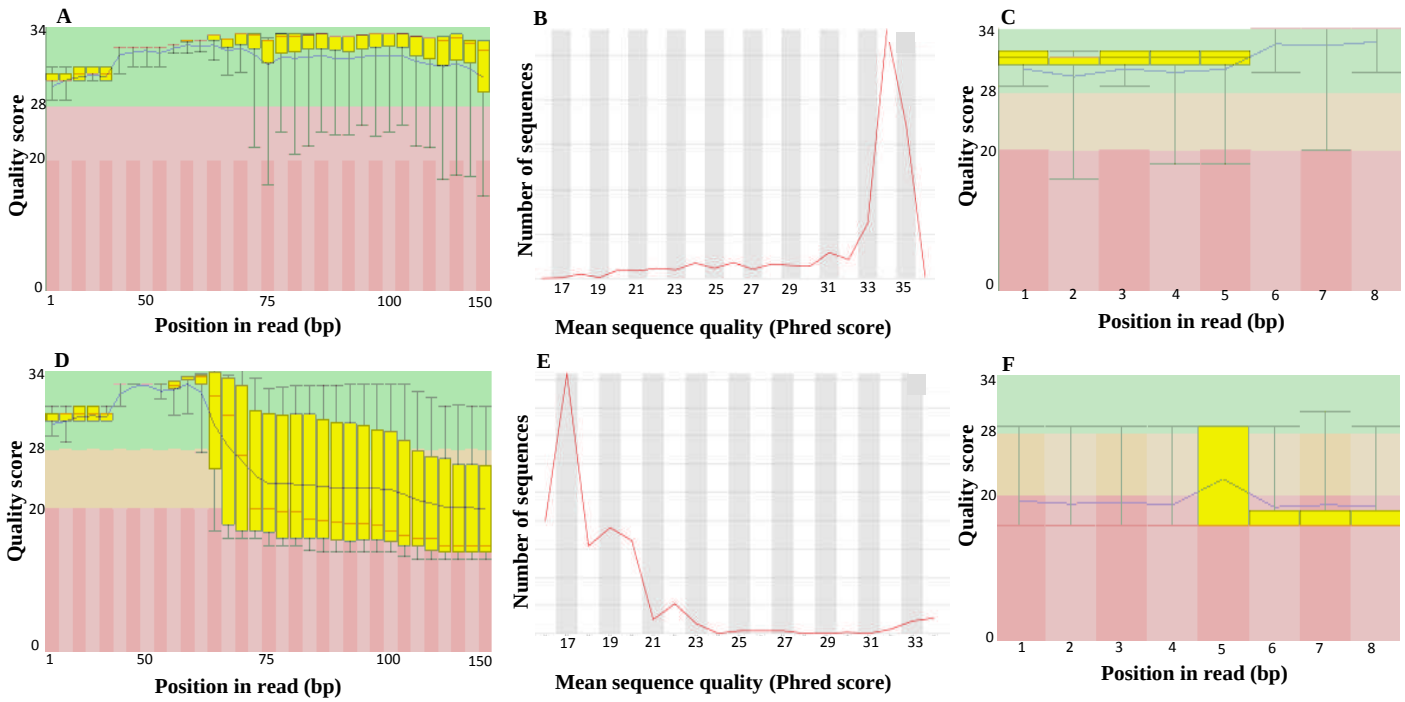

A
Quality score
34
28
20
0
1
50
75
100
150
Position in read (bp)
B
Number of sequences
17
19
21
23
25
27
29
31
33
35
Mean sequence quality (Phred score)
C
Quality score
34
28
20
0
1
2
3
4
5
6
7
8
Position in read (bp)
D
Quality score
34
28
20
0
1
50
75
100
150
Position in read (bp)
E
Number of sequences
17
19
21
23
25
27
29
31
33
Mean sequence quality (Phred score)
F
Quality score
34
28
20
0
1
2
3
4
5
6
7
8
Position in read (bp)

Figure 3.

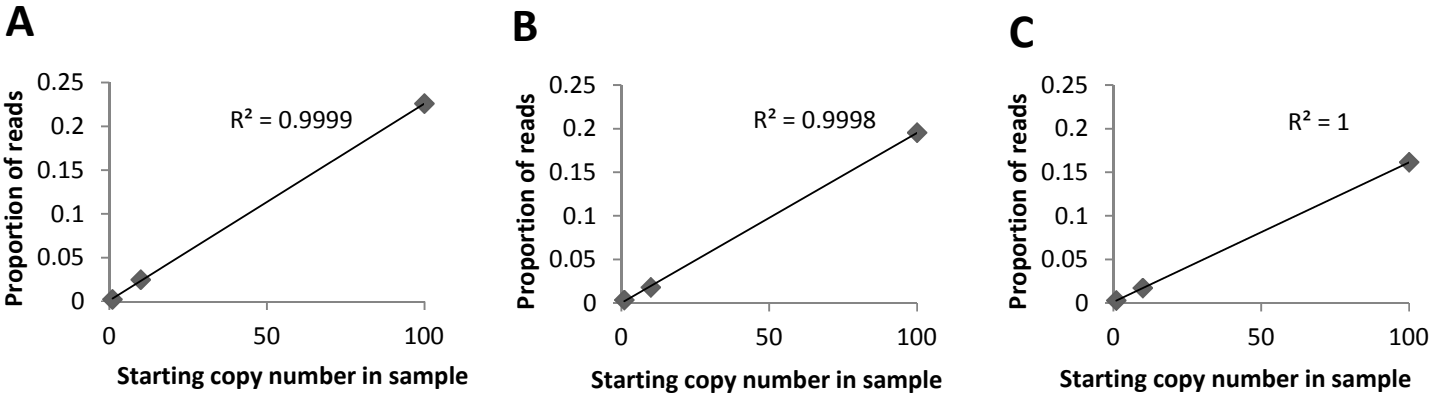

A
B
C
Proportion of reads
0.25
0.2
0.15
0.1
0.05
0
R² = 0.9999
R² = 0.9998
R² = 1
0
50
100
Starting copy number in sample

Figure 4.

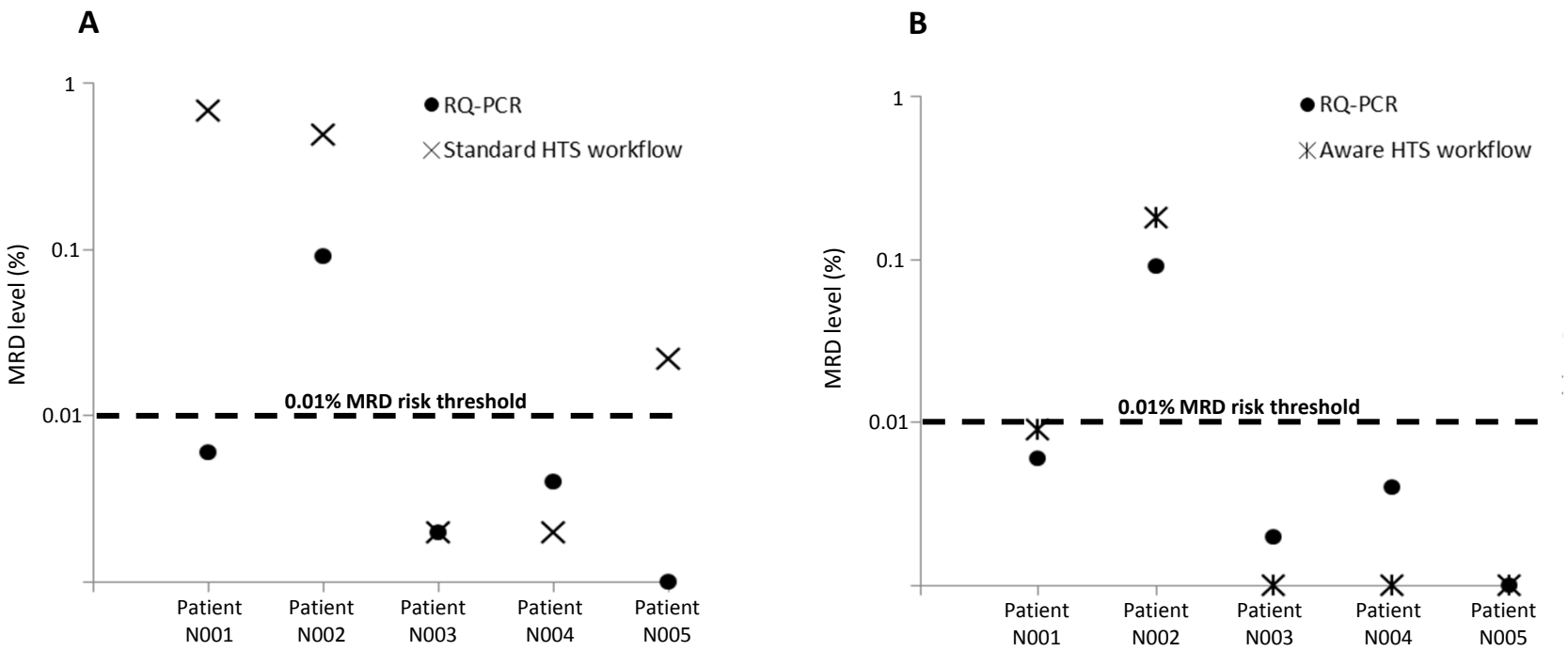

A
1
0.1
0.01
MRD level (%)
RQ-PCR
Standard HTS workflow
0.01% MRD risk threshold
Patient N001
Patient N002
Patient N003
Patient N004
Patient N005
B
1
0.1
0.01
MRD level (%)
RQ-PCR
Aware HTS workflow
0.01% MRD risk threshold
Patient N001
Patient N002
Patient N003
Patient N004
Patient N005

Figure 5.