Published online 10 July 2014

Nucleic Acids Research, 2014, Vol. 42, No. 16 ***e129***
doi: 10.1093/nar/gku607

Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence

Claire T. Deakin[1], Jeffrey J. Deakin[1], Samantha L. Ginn[1], Paul Young[2], David Humphreys[2], Catherine M. Suter[2,3], Ian E. Alexander[1,4,*] and Claus V. Hallwirth[1]

[1]Gene Therapy Research Unit, Children's Medical Research Institute and The Children's Hospital at Westmead, Westmead, New South Wales 2145, Australia, [2]Molecular Genetics Division, Victor Chang Cardiac Research Institute, Sydney, Darlinghurst, New South Wales 2010, Australia, [3]Faculty of Medicine, University of New South Wales, Kensington, New South Wales 2052, Australia and [4]Discipline of Paediatrics and Child Health, The Children's Hospital at Westmead Clinical School, The University of Sydney, Westmead, New South Wales 2145, Australia

Received August 14, 2013; Revised June 10, 2014; Accepted June 24, 2014

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

ABSTRACT

Barcoded vectors are promising tools for investigating clonal diversity and dynamics in hematopoietic gene therapy. Analysis of clones marked with barcoded vectors requires accurate identification of potentially large numbers of individually rare barcodes, when the exact number, sequence identity and abundance are unknown. This is an inherently challenging application, and the feasibility of using contemporary next-generation sequencing technologies is unresolved. To explore this potential application empirically, without prior assumptions, we sequenced barcode libraries of known complexity. Libraries containing 1, 10 and 100 Sanger-sequenced barcodes were sequenced using an Illumina platform, with a 100-barcode library also sequenced using a SOLiD platform. Libraries containing 1 and 10 barcodes were distinguished from false barcodes generated by sequencing error by a several log-fold difference in abundance. In 100-barcode libraries, however, expected and false barcodes overlapped and could not be resolved by bioinformatic filtering and clustering strategies. In independent sequencing runs multiple false-positive barcodes appeared to be represented at higher abundance than known barcodes, despite their confirmed absence from the original library. Such errors, which potentially impact barcoding studies in an application-dependent manner, are consistent with the existence of both stochastic and systematic error, the mechanism of which is yet to be fully resolved.

INTRODUCTION

Retroviral vectors, such as gammaretroviral and lentiviral vectors, have demonstrated great therapeutic potential, particularly for gene therapy applications targeting the hematopoietic compartment. Therapeutic efficacy following retroviral gene delivery to hematopoietic progenitor cells (HPCs) has been reported following trials of gene therapy for several genetic diseases (1–12), leukemia (13) and attenuation of graft-versus-host disease (14). Analyses of vector integration sites (ISs), which uniquely tag individual gene-marked HPC clones, are yielding important insights into clonal complexity, clonal dynamics and genotoxicity following gene therapy. For example, analysis of samples taken 12–102 months post-transplant from eight patients treated in the groundbreaking French SCID-X1 trial showed that diversity of reconstituted T cells correlated positively with the dose of genetically modified HPCs received by each patient (15). Additionally, the proportion of genetically modified HPCs that contributed to long-term hematopoiesis was estimated to be 1%. In the same and subsequent trials involving other disease indications, IS analysis has also been successfully used to investigate adverse events including leukemia, myelodysplasia and non-malignant clonal expansions (16–19). The underlying mechanism proved to be insertional mutagenesis and is now recognized as an important genotoxic risk associated with gene therapy applications using integrating vector systems. While indispensable for investigating the mechanism underlying the above adverse events, IS analysis has a number of limitations when used to assess clonal dynamics, including early and reliable detection of potentially pathological clonal expansions. These limitations include methodological complexity and, with the most widely used protocols involving use of both restriction endonucleases and extensive rounds

*To whom correspondence should be addressed. Tel: +61 2 9845 3071; Fax: +61 2 9845 1317; Email: ian.alexander@health.nsw.gov.au

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

of polymerase chain reaction (PCR), the risk of detection biases that can reduce sensitivity and even preclude detection of certain clones (20). Despite efforts to address these limitations (20–24), there remains considerable impetus for the development of alternative methods with improved sensitivity and greater quantitative potential.

Barcoded vectors, containing random nucleotide (nt) sequences at defined positions, are a conceptually attractive alternative to IS analysis. Individual HPCs would be uniquely tagged provided the barcoded vector stock has sufficiently high complexity. Such an approach could offer more reliable quantitation of clonal contributions if minimal PCR cycles are used to amplify the barcode from the genomic DNA, as well as methodological simplicity. Given that doses in excess of 10^6 transduced HPCs per kg of body weight have been used in hematopoietic gene therapy trials (2,4,6–10), an ideal barcode library may need to contain up to 10^8 different barcodes to ensure HPC clones are uniquely tagged. Analyzing the diversity of such a highly complex barcode library would require the ability to accurately identify large numbers of unique barcode variants of unknown sequence, individually present at low frequency.

The capacity of next-generation sequencing (NGS) to analyze tens to hundreds of millions of short sequence reads raises the possibility of identifying and possibly quantifying very large numbers of barcode variants recovered from genomic DNA extracted from clinical samples. The suitability of existing NGS technologies for this extremely demanding application is yet to be resolved. Current NGS technologies have higher error rates than traditional Sanger sequencing (25,26), and each of the platforms has different error profiles (27,28). Although several analyses of barcodes amplified from integrated retroviral vectors have been reported (29–36), at present it is unknown to what extent sequencing error might impact on the analysis of complex barcoded libraries, and whether there is a limit to the degree of complexity that can be reliably resolved using contemporary NGS technologies. To address these questions empirically, we amplified barcodes of known sequence identity within mixtures of low to moderate complexity using minimal PCR cycles, and sequenced those barcodes using Illumina and SOLiD platforms. Our analysis of these mixtures enabled evaluation of the effect of analytical strategies for reducing background caused by error, the feasibility of setting frequency-based cut-offs for eliminating background, the potential pitfalls that may be encountered when analyzing complex libraries and the extent of contribution to error from PCR and sequencing.

MATERIALS AND METHODS

Barcode design and construction of complex barcoded plasmid libraries

A primer extension method was developed to construct platform-specific double-stranded barcode inserts for cloning into the NsiI site of a previously described lentiviral construct, pEF1α.γc (37), which is based on pRRLsin.cPPT.hCMV.EGFP.WPRE (38) and wherein expression of the common gamma chain (γc) is under the transcriptional control of a 1177-bp human elongation factor 1α (EF1α) promoter-enhancer fragment (Figure 1A). Oligonucleotides were synthesized to contain random nucleotides at defined positions and adaptor sequences for either the Illumina or SOLiD platforms (Supplementary Table S1). Annealing of either primer 5'-[phos]GGCACCCGTGCAC for the Illumina-compatible barcode or primer 5'-[phos]GCTGCTGTACGGCCAAGGCG for the SOLiD-compatible barcode produced an NsiI-compatible end at one end of the barcode insert. The complementary strands of both barcode inserts were synthesized using the 5' → 3' exo$^-$ Klenow Fragment (New England Biolabs) and an NsiI-compatible end was generated at the other end of the barcode insert by cleavage with PstI (New England Biolabs). After ligation of the insert with NsiI-linearized pEF1α.γc, the NsiI site was not reconstituted, which enabled digestion of the ligation product with NsiI to eliminate vector molecules that re-ligated without the barcode insert. Electrocompetent SURE cells (Agilent Technologies) were transformed with the ligation products to produce highly complex Illumina-compatible and SOLiD-compatible barcoded plasmid libraries, with complexities of ∼15 million and 1.8 million, respectively.

Production of defined barcode libraries

From the Illumina-compatible and SOLiD-compatible plasmid libraries, individual plasmids containing 119 and 100 unique barcodes, respectively, were isolated, quantified using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific), and Sanger-sequenced using an AB 3730xl instrument (Australian Genome Research Facility). For all isolated plasmids, concentrations ranged from 36.3 to 235.7 ng/μl. Barcode libraries of defined complexity comprising known sequence identities were produced by mixing the plasmids containing these sequenced barcodes in equimolar proportions. For the Illumina-compatible barcode, plasmids containing unique and defined barcode sequences were mixed to provide libraries containing 10 known and 100 known barcode sequences, the '10-barcode' and '100-barcode' libraries. The 10-barcode and 100-barcode libraries contained six barcodes with the same sequence identities; the sequence comprising the single barcode was also represented within the 100-barcode library (Supplementary Tables S2 and S3). A SOLiD-compatible 100-barcode library was prepared in a similar manner. A single pipette was used during plasmid mixture preparations to minimize pipetting error. The two plasmid mixtures of 100 Illumina-compatible and SOLiD-compatible barcodes contained different barcode sequences, because they were composed of individual plasmids selected at random from the two complex plasmid libraries. The need to incorporate platform-specific adaptor sequences into the barcode inserts necessitated the preparation of separate complex libraries for each platform.

Preparation and NGS of barcode amplicons

The barcode regions in each of the Illumina-compatible and SOLiD-compatible defined libraries were flanked by part of the platform-specific adaptor sequences required for cap-

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

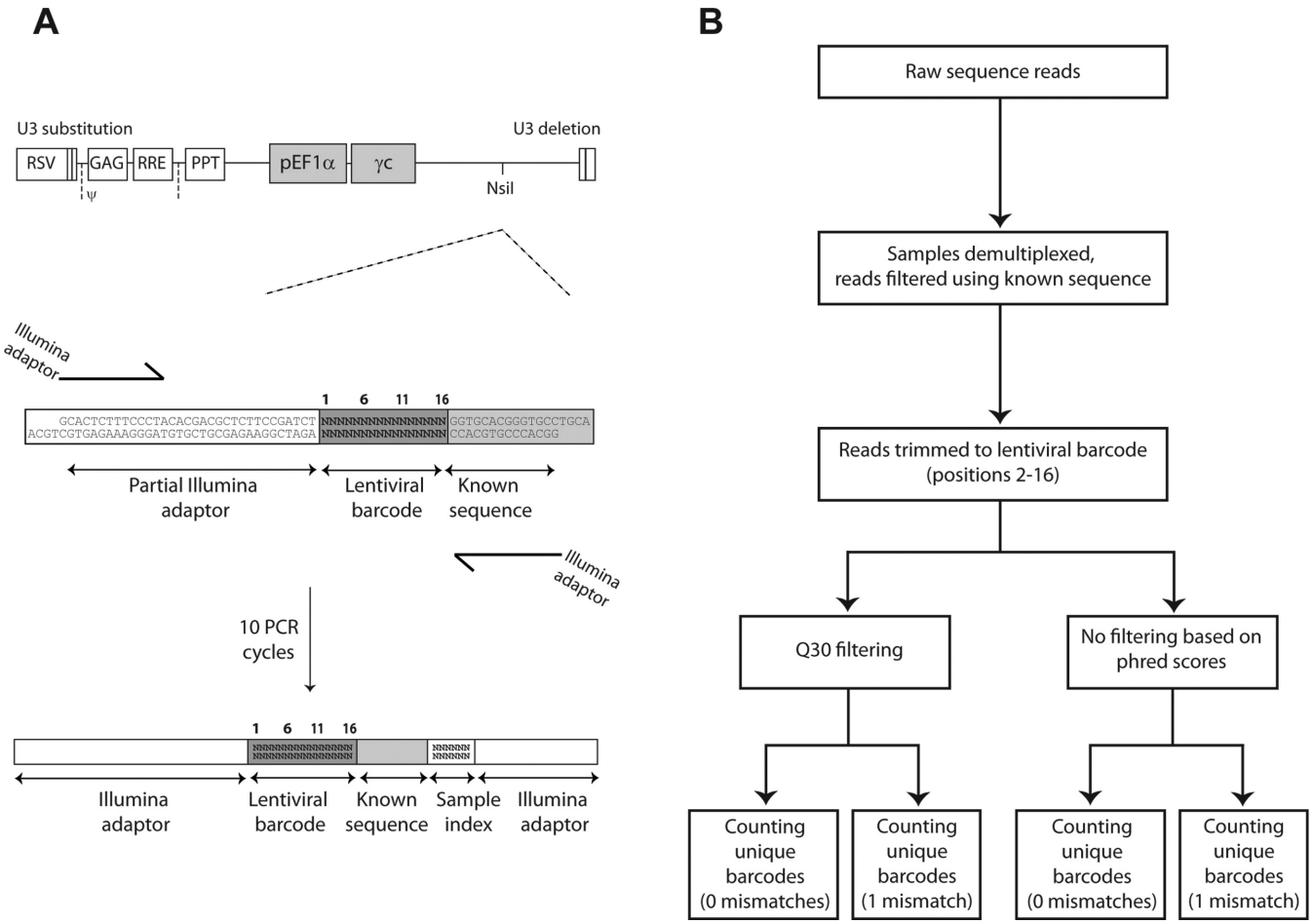

Figure 1. Experimental design and analytical workflow for analysis of the Illumina-compatible barcode. (**A**) Structure and sequence of the Illumina-compatible barcode insert cloned into the NsiI site of the pEF1α.γc lentiviral construct. The insert contained a PstI site, 32 bp of the Illumina adaptor sequence, a 16-bp random sequence that functioned as the lentiviral barcode and an 18-bp known sequence. Numbers indicate the position of every fifth random nucleotide in the barcode. The SOLiD-compatible barcode followed a similar configuration, with the insert containing a PstI site, 23 bp of the P1-T adaptor, a 15-bp random sequence for the lentiviral barcode and the internal adaptor. For both barcode configurations, the barcode regions were amplified with 10 PCR cycles using primers that introduced the adaptor sequences required for the Illumina or SOLiD platforms. (**B**) Strategy for analyzing sequence data for the Illumina-compatible barcode. Raw sequence reads were filtered using the known sequence immediately following the barcode at positions 17–30 to eliminate indel errors. The lentiviral barcode was trimmed to positions 2–16 to avoid errors at position 1. The number of unique barcode sequences was counted with and without phred score filtering (Q30), and with and without allowing one mismatch. For the SOLiD-compatible barcode, raw sequence reads were filtered using 10 internal adaptor sequences and the number of unique barcode sequences were counted with and without allowing one mismatch.

ture and sequencing on either the Illumina or SOLiD platforms (Figure 1A). Barcode samples were therefore PCR-amplified using primers that introduced the remaining portions of the adaptor sequences as well as a sample index (Supplementary Table S1). High fidelity Phusion polymerase (Thermo Fisher Scientific) and only 10 PCR cycles (95°C for 30 s, 55°C for 30 s and 72°C for 30 s) were used for amplification, to minimize the potential for PCR bias and polymerase error. Amplicons were gel-purified and quantified using the NanoDrop 1000. Amplicon samples of the single barcode, and the 10- and 100-barcode libraries were mixed to contain ∼100 000 copies of each barcode amplicon. Sequencing of Illumina-compatible barcodes was conducted on a HiSeq 2000 instrument using 50 base single-end reads (Beijing Genomics Institute), with 22 057 163 reads dedicated to these samples. The same plasmid mixture that constituted the Illumina-compatible100-barcode library was amplified for sequencing a second time on the HiSeq 2000, producing 82 508 636 reads. Small-scale sequencing of SOLiD-compatible barcodes (1 219 079 reads) was conducted on a 5500xl instrument (Victor Chang Cardiac Research Institute). For both libraries sequence heterogeneity was ensured at the proximal ends of sequence reads, since they commenced with the 16- and 15-nt barcode sequences of the Illumina-compatible and SOLiD-compatible libraries, respectively. To further maximize sequence heterogeneity in the first Illumina sequencing run, the defined libraries were spiked into a complex background that comprised 90% of sequence reads. Thus the single barcode, 10- and 100-barcode libraries comprised 0.09%, 0.9% and 9% of sequence reads, respectively. To examine the potential for the 10-cycle PCR to influence the relative abundance of barcodes, selected barcode amplicons were mixed in equimolar proportions after PCR amplification and gel-purification. Samples derived from corresponding pre- and post-amplification mixing were sequenced in a third inde-

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

pendent sequencing run on the HiSeq 2000, with 173 560 reads dedicated to these samples.

Data filtering, analysis of unique barcodes and analysis of a constant region flanking the barcode

Raw sequence reads obtained from the Illumina HiSeq 2000 were initially filtered for the known sequence, 'GGTGCACGGGTGCC', at positions 17–30 (Figure 1B). This facilitated elimination of errors caused by nucleotide insertions or deletions. Subsequent analyses of the processed reads were conducted using a combination of standard UNIX tools for string manipulation, the MySQL (Version 5.1.47) relational database for sequence counting and customized Perl scripts for data manipulation, filtering, clustering and error analysis (scripts provided as Supplementary Methods). Reads were trimmed to positions 2–16 of the barcode using barcode-parse.pl, since previous analyses of the distribution of error frequencies at different nucleotide positions during preliminary sequencing runs had indicated that the first position of an Illumina read can be highly error-prone (data not shown), a phenomenon also reported by others (27). Reads wherein any of the barcode positions had a Phred score below 30 were filtered out using barcode-filter.pl to produce Q30-filtered data. Unique barcode sequences were counted and then listed in order of decreasing abundance using a simple MySQL query. The clustering program, cluster.pl, was designed to assume no prior knowledge of real versus artifactual barcodes and processed barcodes in a hierarchical fashion using the correct order produced by MySQL. Briefly, all detected sequences were compared with the first-most abundant barcode and their counts were added to those of the first-most abundant barcode if they differed by one position, which was determined by calculating the Hamming distance of each sequence relative to the first-most abundant barcode. The process was repeated for the second-most abundant barcode and then the third-most abundant barcode, etc., until all remaining barcodes had been processed. Raw sequence reads obtained from the SOLiD 5500xl were processed similarly, with filtering based on the presence of the first 10 nt of the internal adaptor sequence, 'ACGCCTTGGC', at positions 16–25, followed by one-mismatch clustering. The known sequence, 'GGTGCACGGGTGCC', at positions 17–30 of the Illumina-compatible barcode (Figure 1A) was analyzed similarly. Raw sequence reads from the first and second Illumina sequencing runs were filtered based on the presence of the expected index sequences at the expected positions, 31–36. Reads were then trimmed to positions 17–30 and Q30-filtered using barcode-filter.pl. The number of unique sequences was counted using MySQL, and sequences that differed by one mismatch were clustered using cluster.pl. The potential misassignment of reads to another sample (attributable to incorrect assignment of Illumina index sequences) was assessed for the three libraries sequenced in the first sequencing run. Using agrep and allowing two mismatches, Q30-filtered reads were screened for barcode sequences belonging to the other respective samples. Barcode sequences known to be contained within more than one library were excluded from this analysis.

The clustering program was also used to establish the maximum number of mismatches that could be permitted before each of the known sequences in the 100-barcode libraries could no longer be unambiguously identified. Although up to five mismatches could theoretically be tolerated, in practice allowing two to four mismatches did not eliminate high-frequency false-positive barcodes or distinguish expected barcodes from background. Within the plasmid mixture of 100 Illumina-compatible barcodes, 18 plasmids contained at least two barcodes. For these barcodes the relative abundance of the apparently least abundant barcode was compared to that of the other barcode(s) present on the same plasmid molecule, following sequencing in the first and second Illumina runs, Q30 filtering and one-mismatch clustering.

Analyses of empirical error rates, one-mismatch errors and barcode sequence characteristics

Empirical error rates were assessed using mismatch-barcode.pl, which compared each position in the barcode region to the expected nucleotide for that position across all reads of the single barcode. Analysis of one-mismatch errors was performed using error-analysis.pl, which compared the one-mismatch error sequences identified by cluster.pl to the expected barcode sequences from which they differed by one mismatch. The first 89 records from the cluster.pl output were used, since these contained expected barcode sequences. This method assumed that the sequences that differed from expected barcodes by one mismatch were generated by errors at the mismatch position. While this analysis provided insight into the type and location of stochastic single nucleotide substitution-like errors, it was uninformative with regard to the systematic errors that resulted in high-frequency false barcodes that appeared to contain six or more mismatches. GC content for each of the 100 known barcode sequences was calculated using gatc-string.pl. Minimum Gibbs Free Energy (MFE) values for the 100 known barcode sequences and whole barcode amplicons were calculated using UNAfold (39). Putatively false barcodes that were detected within the top 120 unique barcodes for the Illumina-compatible barcode library were compared to each of the 100 expected barcodes using Hamming distances calculated by 100-noise-hamming-distance.pl. The maximum length of perfect homology between high-frequency false barcodes and the 100 known barcodes was calculated using compare-false-to-known.pl. Multiple sequence alignments and analyses of conserved regions were conducted using BioEdit (Version 7) (40).

PCR assays to determine individual barcode representation within the 100-barcode library

The presence of individual barcode sequences within the Illumina-compatible 100-barcode plasmid mixture was assessed using semi-quantitative PCR. Each PCR assay used a primer designed to anneal specifically to the barcode sequence and a vector-binding primer designed to produce an ~120-bp product (Supplementary Table S1). The 3' end of the barcode-specific primers contained 11 nt that specifically bound to the barcode, such that the remaining 5 nt of

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

the barcode sequence served as template for amplification. These 5 nt functioned as a 'signature' for each barcode sequence and were used to confirm the specificity of amplification when PCR products were Sanger-sequenced. Although each PCR assay was set up as a quantitative PCR (qPCR), the constraints on primer placement that were necessary to ensure specific amplification of barcodes were not compatible with optimal qPCR primer design. Consequently, the assays were used to assign the absence or presence of specific templates within their specific detection limits. For each PCR assay, a 1×10^{10}-copy standard of the plasmid containing the barcode of interest was prepared and then serially diluted 1 in 10 into the pEF1α.γc construct without the barcode insert, such that 1×10^{5}-copy to 1×10^{9}-copy standards were produced for constructing standard curves. Ten million-copy samples were prepared for each of the samples that were analyzed, which included the plasmid mixture comprising the 100-barcode library, a 10-barcode plasmid mixture including the barcode of interest, the plasmid containing the barcode of interest and three plasmids containing different barcodes. Reactions used the SYBR Green JumpStart Taq ReadyMix (Sigma-Aldrich) and were analyzed using a Rotor-Gene 6000 real-time PCR detection system (Corbett). Cycles consisted of an initial denaturing step (95°C for 10 min), followed by 25 cycles of template denaturing (95°C for 20 s) and primer annealing and extension (72°C for 20 s) and a final extension step (72°C for 7 min). The specificity of barcode amplification was confirmed by Sanger-sequencing of the PCR product that was produced when the 100-barcode library served as template for amplification, for those samples that yielded an amplification product.

In order to estimate the limit of sensitivity of this approach for detecting specific barcodes, a barcode sequence that was absent from the 100-barcode library was spiked into the 100-barcode library in known proportions (0.1%, 0.5%, 0.75%, 1%, 1.25%, 2.5%, 5%, 10% and 20%) and analyzed by semi-quantitative PCR using a recessed primer specific for that barcode (Supplementary Table S1). The limit of sensitivity of this approach for detecting specific barcodes was thus established as ≥0.1%. The specificity of amplification of the spiked-in barcode sequence was validated by Sanger sequencing of the products yielded when the mixtures containing 0.1% and 1% of the spiked-in barcode were used as template. Additionally, end-point PCR was performed to detect the presence of selected barcodes. A similar recessed primer design was employed for the barcode-specific primers and two different vector-binding primers were used to accommodate either forward or reverse orientations of the barcode insert (Supplementary Table S1). Cycling conditions were identical to those used for the semi-quantitative PCR. End-point PCR was employed to evaluate the presence or absence of certain barcodes in four different samples of the 100-barcode library after the 10 amplification cycles. The samples analyzed included the two post-amplification samples that were used in the Illumina HiSeq 2000 sequencing runs, of which insufficient quantities remained for analysis using semi-quantitative PCR.

Statistical analyses

Comparisons between observed and expected proportions of errors at barcode positions and each of the possible types of substitution errors were made using χ^2 tests (GraphPad Prism Version 5). Comparisons of the observed and expected abundance of barcodes mixed after 10 PCR cycles were also performed using χ^2 tests. Comparisons of the maximum length of perfect homology between certain barcodes and the 100 known barcodes were performed using a *t*-test (GraphPad Prism). Correlation between the detected abundance of barcodes in the 100-barcode library during the first and second sequencing runs was calculated using Pearson R (GraphPad Prism).

RESULTS

Sequencing of defined barcodes reveals distinct categories of error

Barcoded plasmid libraries were constructed, and the barcode regions were amplified from a single known barcode and the defined 10- and 100-barcode libraries using 10 PCR cycles, for sequencing on the Illumina platform (Figure 1A). Additionally, a separate defined 100-barcode library was sequenced on the SOLiD platform. For each of these defined samples, raw sequence reads were processed and the number of unique barcode sequences was analyzed (Figure 1B).

One-barcode sample. For the sample containing a single barcode, 92.66% of Q30-filtered sequence reads were accounted for by the expected sequence when one mismatch was permitted (Table 1). The remaining reads, however, detected 8099 unique false-positive barcodes in this sample known to contain only a single barcode. Misassignment of reads originating from the 10-barcode and 100-barcode libraries accounted for 0.013% of reads (or 0.17% of observed error). The distribution of the relative frequency of the 500 most abundant barcodes detected in the one-barcode sample, however, indicated that the expected barcode had greater than four log-fold higher abundance than background (Figure 2A and F). Background could thus be eliminated using a frequency-based cut-off. The average empirical error rate across the barcode positions for the single barcode was calculated as 4.62% (SD = 0.15%) after Q30 filtering and 5.55% (SD = 0.30%) before Q30 filtering. Empirical error rates were not calculated for the 10- and 100-barcode libraries because these samples did not have a unique reference sequence.

10-barcode library. The number and proportion of background sequences were lower for the 10-barcode library compared to the one-barcode sample. The 10 expected barcodes accounted for 99.92% of Q30-filtered sequence reads when one mismatch was tolerated (Table 1). Incorrect assignment of reads of the single barcode and barcodes in the 100-barcode library contributed 0.016% of sequence reads (19.0% of total error). Although 520 putatively false barcodes were detected, the distribution of the relative frequency of the 500 most abundant barcodes indicated that the 10 expected barcodes had an ∼3 log-fold higher abundance than background (Figure 2B and F). Interestingly,

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

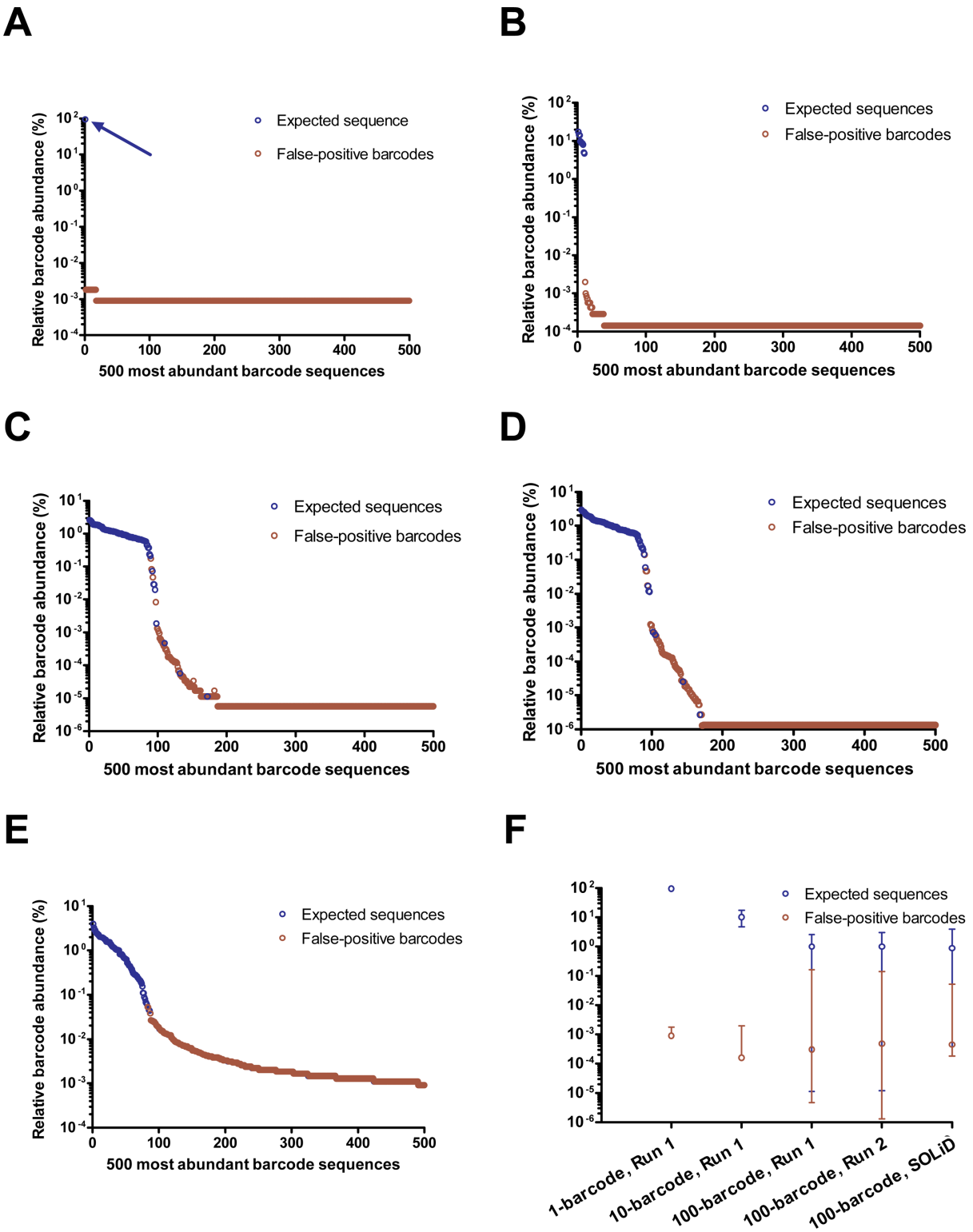

Figure 2. Distribution of the relative abundance of the 500 most abundant barcode sequences detected following analysis of the defined barcode libraries using different sequencing platforms. Libraries containing (**A**) 1, (**B**) 10 and (**C**) 100 defined Illumina-compatible barcode(s) sequenced using the first sequencing run. For the 100-barcode library, the first 89 most abundant barcodes matched expected sequences, and a point of inflection in the distribution of the relative frequencies of barcodes occurred at the 82nd-most abundant barcode. Six putatively false barcodes that were not in the 100-barcode library were detected within the top 100. (**D**) Library containing the same 100 defined Illumina-compatible barcodes sequenced using the second sequencing run after an independent amplification. Seven putatively false barcodes were detected within the top 100. A point of inflection occurred at the 79th-most abundant barcode and again, the first 89 most abundant barcodes matched expected sequences. (**E**) Library containing 100 defined SOLiD-compatible barcodes. The first 82 most abundant barcodes matched expected sequences; however, 13 putatively false barcodes were detected in the top 100. (**F**) Mean and range of relative abundances of expected and false barcodes, for each sample.

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

Table 1. Effect of analytical strategies on the reduction of background error

		Illumina HiSeq 2000				SOLiD 5500xl
		1-barcode	10-barcode	100-barcode	100-barcode	100-barcode
	Property	(Run 1)	(Run 1)	(Run 1)	(Run 2)	
Analysis	Fold coverage of each barcode	134 364×	81 714×	211 057×	825 086×	12 191×
Counting unique	Processed sequence reads	134 364	817 135	21 105 664	82 508 636	1 219 079
	Total unique barcodes detected	9691	1073	7977	14 672	112 624
	Processed reads accounted for by expected sequence(s)	91.45%	98.02%	97.59%	97.73%	55.86%
Q30 filtering, counting unique	Processed sequence reads[a]	111 766 (16.8%)	705 473 (13.7%)	17 694 272 (16.2%)	75 076 957 (9.0%)	546 409 (55.2%)
	Total unique barcodes detected	8219	819	3638	5241	7068
	Processed reads accounted for by expected sequence(s)	92.48%	99.55%	99.32%	99.42%	96.86%
Q30 filtering, clustering 1 mismatch, counting unique	Processed sequence reads	111 766	705 473	17 694 272	75 076 957	546 409
	Total unique barcodes detected	8100	530	687	1014	6963
	Processed reads accounted for by expected sequence(s)	92.66%	99.92%	99.67%	99.74%	97.33%
Q30 filtering, clustering 1 mismatch, excluding error-prone positions, counting unique	Processed sequence reads	113 198	717 794	18 109 912	75 532 263	n.a.[b]
	Total unique barcodes detected	8351	545	706	902	n.a.
	Processed reads accounted for by expected sequence(s)	92.58%	99.92	99.67%	99.74%	n.a.

[a]Proportion of processed sequence reads eliminated during Q30 filtering given in parentheses.
[b]n.a., analysis not performed.

the 10 known barcodes were not detected in the expected equal proportions (Supplementary Table S2).

100-barcode library. For the 100-barcode library, while the number and proportion of background sequences were also low, expected barcodes overlapped with background (Figure 2F). The 100 expected barcode sequences accounted for 99.99% of Q30-filtered sequence reads when one mismatch was allowed, with the remaining background comprising 677 putatively false barcodes (Table 1). Misassigned reads of barcodes from the 10-barcode library accounted for 0.00061% of reads (0.19% of total error). The distribution of the relative frequency of the 500 most abundant barcode sequences displayed no clear distinction between the 100 expected barcodes and background, with some expected barcodes being detected at lower abundance than putatively false barcodes (Figure 2C, Supplementary Table S3). Three of the expected barcodes were detected outside of the 100 most abundant sequences, and a further three expected barcodes were undetected. When the same plasmid mixture of 100 barcodes was sequenced using 4.2-fold higher coverage in an independent run following an independent 10-cycle PCR, a similar pattern was observed. The 100 expected barcodes accounted for 99.74% of filtered sequence reads after Q30 filtering and clustering, and 914 putatively false barcodes were detected (Table 1). Although baseline levels of background were lower compared to the previous sequencing run of the same sample, the distribution of the 500 most abundant barcodes displayed a similar overlap between the 100 expected barcodes and background (Figure 2D and F, Supplementary Table S3). Five of the ex-

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

pected barcodes were detected outside of the top 100, and a further two expected sequences were not detected. A barcode library containing 100 known sequences could therefore not be fully resolved from background using the Illumina platform.

SOLiD-compatible library. A separate 100-barcode library was sequenced using the SOLiD platform. This was a preliminary analysis to explore whether an independent technology was more suitable for analyzing complex barcode libraries, given the potential advantages of di-base interrogation and color-space mapping. While the 100 expected barcodes accounted for 97.33% of filtered sequence reads, 6863 putatively false barcodes were detected (Table 1). The distribution of the 500 most abundant sequences indicated that the expected sequences could not be distinguished from background (Figure 2E and F; Supplementary Table S4). Eleven of the expected barcodes were detected outside of the top 100 and a further two expected barcodes were not detected.

Analysis of a known sequence outside the barcode. The number and abundance of unique sequences detected were analyzed for a constant region downstream of the barcode that was introduced by the primers used to amplify the barcode and also sequenced using the Illumina platform (Figure 1A). The expected sequence accounted for 99.99% and 99.999% of sequence reads after Q30 filtering and clustering for the first and second Illumina sequencing runs, respectively, with totals of 146 and 141 unique sequences detected for the first and second runs, respectively (Supplementary Table S5). The overall proportion of error for this known sequence was lower than that for the barcode region in each of the samples. Errors detected in the barcode region represented the combination of errors generated during the 10-cycle PCR and sequencing. Therefore, further analyses sought to attribute the source of errors detected in the barcode region to either PCR or sequencing for each of the categories of error observed, namely, discrepancies of relative barcode abundance, loss of barcodes, generation of false barcodes and random errors.

Discrepancies of relative barcode abundance arise during sequencing

Analyses of the Illumina-compatible 10- and 100-barcode libraries revealed a discrepancy between the expected equal abundance of each of the known barcode sequences and an observed uneven abundance. For the 10-barcode library, where each barcode was expected at 10% abundance, there was a 3.7-fold discrepancy between the most and least frequently detected known barcodes (Supplementary Table S2). The observed uneven distribution of apparent abundance for each of the known barcodes in the 100-barcode library, expected at 1%, followed a pattern that was consistent overall, but non-identical for both sequencing runs (Figure 3A). The most frequently detected known barcodes were detected at 2.59% and 2.99% abundance in both sequencing runs, while the 89th-most abundant barcodes were detected at 0.21% and 0.15% abundance by the first and

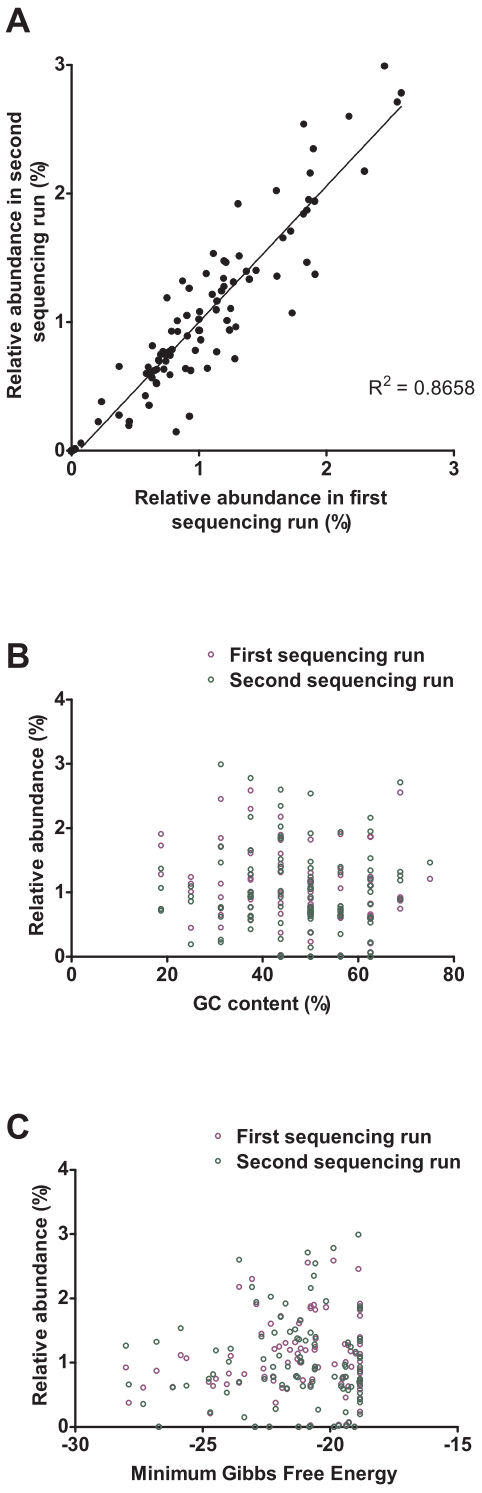

Figure 3. Analysis of the relative abundance, GC content and likelihood of secondary structure formation for each of the 100 expected Illumina-compatible barcode sequences. (**A**) Relative abundance of the 100 expected barcode sequences, as detected during the first and second sequencing runs using the Illumina HiSeq 2000 (Pearson $r(98) = 0.93$, $p<0.0001$). (**B**) Distribution of the relative abundance of each barcode sequence as a function of the percentage GC content of that sequence. (**C**) Distribution of the relative abundance of each barcode sequence as a function of the MFE value calculated for that sequence. MFE values provide an estimate of the likelihood of secondary structure formation, with lower values associated with a higher likelihood.

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

second sequencing runs, respectively. The sequence identities of the three most abundant barcodes and the three undetected barcodes were the same between sequencing runs (Supplementary Table S3). The relative abundance of each of the 100 barcode sequences did not correlate with GC content or MFE values, as a prediction of the likelihood of secondary structure formation (Figure 3B and C). Variance in GC content was high for both high-abundance and low-abundance barcodes, and the amplitude in MFE values was small overall. Multiple sequence alignments and analyses of conserved regions failed to reveal any common motifs among the 10 most abundant and 10 least abundant barcodes.

Since some of the plasmids used to prepare the plasmid mixture of 100 barcodes contained multiple barcode inserts, the relative abundance of barcodes present on the same molecule was compared. For plasmids containing two or more barcodes, 11 out of 23 and 13 out of 23 barcodes were detected at greater than 1.5-fold higher abundance than the less abundant barcode in the first and second sequencing runs, respectively. The plasmid molecule that contained one of the undetected barcodes contained a second barcode, which was detected at 0.68% and 0.71% relative abundance in the first and second sequencing runs, respectively. To map the source of consistent over- and under-representation of certain barcodes, two mixtures of barcodes were prepared so as to be represented at equimolar proportions after PCR and before sequencing. The observed abundance of barcodes in each of these mixtures was still unequal after an independent Illumina sequencing run and differed significantly from the expected equal abundance ($p < 0.0001$; Supplementary Table S6).

Loss of barcodes is attributable to PCR, not sequencing

The same three barcodes were undetected (or detected below background) in two Illumina sequencing runs (Supplementary Table S3). PCR analysis detected these barcodes in the original pre-amplification plasmid mixture, but not in the post-amplification barcode mixtures. Therefore, the loss of barcodes was attributable to the 10-cycle PCR used to introduce the Illumina adaptor sequences.

Generation of false barcodes during sequencing

Certain unexpected barcodes were consistently detected within the 120 most abundant barcodes for the 100-barcode library in both sequencing runs, despite implementation of background reduction strategies (Table 1). When the false barcodes detected in both sequencing runs were compared to the 100 expected sequences at each position, they differed by a minimum of 6 nt from the expected sequences to which they were most closely related. Six mismatches exceeded the maximum of five mismatches that could be tolerated in retaining lack of ambiguity among the 100 expected sequences. Additionally, the maximum lengths of homology were calculated when the six highest-frequency false barcode sequences, their reverse complements and six randomly generated sequences were compared to each of the 100 expected sequences. There was no evidence of a difference in the maximum lengths of homology for the six false barcodes or their reverse complements relative to the six random sequences ($p = 0.55$ and 1.0, respectively). The high-frequency false barcodes thus bore no recognizable resemblance to the known sequences. The six putatively false barcodes could not be detected by PCR in the original plasmid mixture containing the 100 expected barcodes. For the first two of these barcodes, which were detected at 0.17% and 0.14% abundance, and 0.05% and 0.08% abundance, in the first and second sequencing runs, respectively, the sensitivity of the PCR analysis method used was $\geq 0.01\%$ (Supplementary Figure S1). These two sequences also failed to be detected in the two post-amplification samples of the 100-barcode library that were sequenced, indicating that their apparent presence arose during sequencing and not during the 10 cycles of PCR.

Characterization of error location and substitution type for stochastic errors

The location and substitution type of stochastic errors identified by one-mismatch clustering of Q30-filtered data were characterized. Analyses revealed that the locations of errors were not evenly distributed and differed between sequencing runs (Figure 4A). Elimination of the more error-prone positions resulted in fewer reads being discarded, but overall levels of background remained unchanged (Table 1). Analyses of each of the 12 possible types of substitution-like errors indicated that these were also unevenly distributed (Figure 4B). There was an over-representation of substitutions to G and an under-representation of substitutions to C, both of which were reproducible across both sequencing runs.

DISCUSSION

The growth of clinical trial activity for gene therapy targeting the hematopoietic compartment has spurred the need to develop improved methods for monitoring clonal diversity and size, and particularly greater sensitivity for early detection of potentially pathological clonal expansions. Since the sequence identities of barcodes in a barcode library of the complexity required for clinical use would be unknown, the veracity of barcode variants identified using NGS must be assumed, unless there is a clear distinction between true barcodes and erroneously generated sequences. In this study, the feasibility of vector barcoding coupled with NGS was investigated for potential use in analyzing clonal diversity, using mixtures of defined complexity comprising known barcode sequences. We describe an empirical approach to evaluate the size of a barcode library that can be resolved by NGS, using over 100 000-fold coverage in two independent sequencing runs to analyze the same library of known barcode sequences. In doing so, similar limitations applicable to independent NGS technologies were identified, pertaining to the analyzable degree of complexity, sensitivity and specificity of analysis and assessable clone size.

In the context of our experimental configuration, sequencing error was found to impose an upper limit on the degree of complexity that could be resolved using NGS. Barcode libraries of low complexity, containing just one or 10 unique barcodes, could readily be distinguished from

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

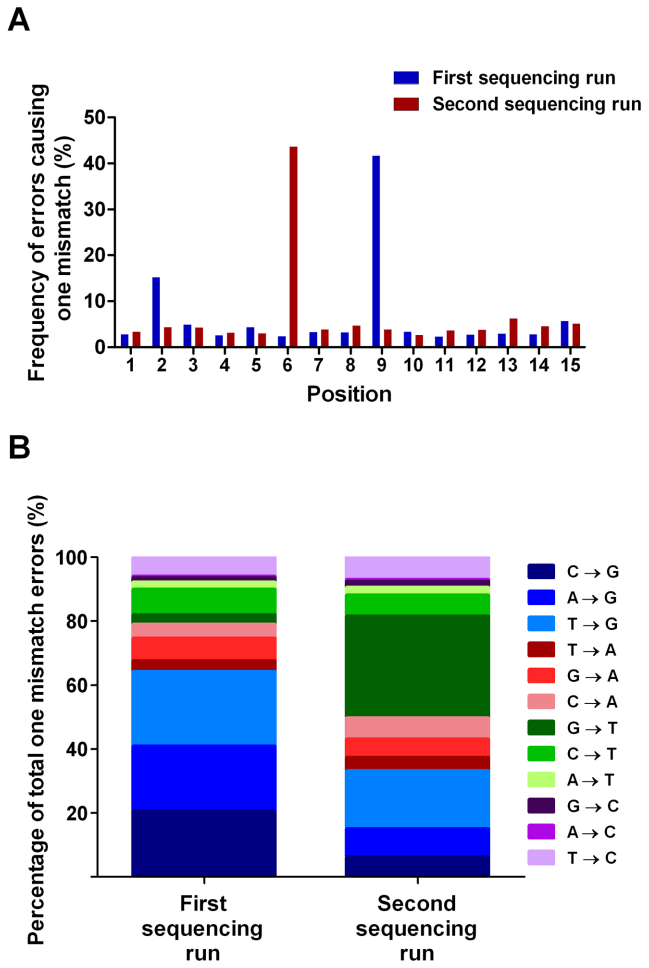

Figure 4. Analysis of the position and substitution-like type of error for all one-mismatch sequence errors for both Illumina HiSeq 2000 sequencing runs. One-mismatch errors were compared to the known barcode sequences from which they were derived. Errors from the first sequencing run represent the sum of one-mismatch errors after Q30 quality filtering for the one-barcode sample and 10- and 100-barcode libraries, although one-mismatch errors from the 100-barcode library comprise 95.3% of all errors. Errors from the second sequencing run represent one-mismatch errors after Q30 quality filtering for the 100-barcode library. (**A**) Distribution of one-mismatch errors across each position of the barcode (positions 2–16 of the sequence reads). This distribution differed significantly from an expected even distribution ($\chi^2 = 30\,064$, df = 14, $p < 0.0001$ for the first sequencing run; $\chi^2 = 90\,717$, df = 14, $p < 0.0001$ for the second sequencing run). (**B**) Distribution of each possible substitution-like error type. This distribution also differed significantly from an expected even distribution ($\chi^2 = 26\,127$, df = 11, $p < 0.0001$ for the first sequencing run; $\chi^2 = 82\,229$, df = 11, $p < 0.0001$ for the second sequencing run). df, degrees of freedom.

background produced by error. However, this was not the case for libraries containing 100 barcodes. While implementation of bioinformatic strategies reduced the total levels of background to below 0.5% of total sequence reads for a library containing 100 barcodes, no cut-off could be established between foreground and background. In the experimental configuration used here, the upper limit on the analyzable degree of complexity thus lies between 10 and 100 known barcodes. It is estimated that a barcode library with complexity of 2500 would be sufficient to mark 90% of transduced HPCs that undergo hematopoiesis in a clinical trial (36). Our findings therefore imply that it may be difficult to use current NGS technologies to fully resolve library complexities in the order of magnitude required for clinical applications. However, this does not negate the use of barcoding in studies that are less sensitive to the level of background comprising individual false positives, such as comparisons of barcode complexity in different treatment groups or cell types.

Our analyses uncovered four categories of error, namely stochastic errors, generation of false barcodes, discrepancies of relative barcode abundance and loss of barcodes. For each of these error categories, we sought to discriminate between PCR and sequencing as the dominant source. We minimized the contribution of polymerase infidelity to stochastic error by using just 10 PCR cycles and the high-fidelity Phusion polymerase, which has a reported error rate of 4.4×10^{-7} (41). At this error rate, 0.0066% of PCR product molecules would have contained errors in the 15 bp barcode sequence. By contrast, NGS technologies are known to have error rates several orders of magnitude higher, ranging from 10^{-3} to 10^{-2}, and greater than 10^{-2} in certain sequence contexts such as motifs with high GC content (42–44). In our analysis of the single barcode, the average error rate was 4.62% for Q30-filtered reads. However, since the overall proportion of background detected for that sample was higher than for the other samples in the same sequencing run, it is likely to overestimate the true error rate for that run. Even at a low sequencing error rate of 10^{-3}, 1.5% of sequence reads would contain errors. Therefore, the stochastic errors contributed to our data by sequencing are expected to outnumber the PCR-generated ones by more than two orders of magnitude. Indeed, errors were also detected within the known sequence outside the barcode, which was introduced by the primers used for amplification and was therefore not prone to polymerase error (although a proportion could represent errors within the primers due to inefficient coupling during oligonucleotide synthesis). The impact of stochastic errors upon the analysis of a complex library can be reduced using analytical strategies such as the clustering approach we used, which in principle is similar to many of the reported error correction algorithms (45–49).

Our study also revealed biases toward certain types of stochastic sequencing error. Many of the stochastic sequencing errors generated on the Illumina platform in this study were single nucleotide substitution-like errors, the majority of which occurred at specific positions in a sequencing run-dependent manner, yet involved specific types of substitutions in a sequencing run-independent manner. The dependence and independence of position effects and substitution effects, respectively, on individual sequencing runs implies that these effects are not related. Such biases in error types underscore the value of including a known barcode sequence, which in this study facilitated development of analytical strategies, including identification of any error-prone positions. The observation of biases toward certain types of stochastic sequencing errors is consistent with other reports (47,50), although the specific biases observed in this study have not previously been reported. For example, analysis of single Sanger-sequenced TCR sequences found most erroneous reads could be accounted for by C to T, G to A, A to G and T to C substitutions (51).

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

Systematic errors, such as those that gave rise to high-frequency false barcodes in our study, have implications distinct from stochastic errors and cannot be computationally curtailed using the clustering-based approach. The same six false barcodes, which were detected in the top 100 barcodes after two independent sequencing runs, shared little homology with any of the expected barcode sequences, suggesting they were unlikely to have arisen from recombination or polymerase strand jumping during PCR. Our analyses of the same two post-amplification 100-barcode library samples that were sequenced after independent amplification eliminated PCR as the source of these errors. Intriguingly, such high-frequency errors entailed the same substitution-like error occurring at the same nucleotide thousands of times. For example, the most frequently detected false barcode was counted 30 529 and 105 067 times after Q30 filtering and clustering of reads generated by the first and second Illumina sequencing runs, respectively. Although misassigned reads of known barcodes did not account for the majority of the error detected in the single barcode, and 10-barcode and 100-barcode libraries sequenced during the first run, it is likely that misassigned reads originating from the complex background into which these samples were spiked would have contributed additional error. This highlights a need for vigilance when using multiplexed index sequences.

Our findings are consistent with a recently reported observation that even three defined barcodes of different abundance could not be resolved from false-positive barcodes in all 28 replicates of a control experiment (29). The observation of this high-abundance false-positive barcode occurred despite the use of a more tolerant clustering approach (applied to fewer variable barcode positions) and additional detection thresholds that were not implemented in our study. Moreover, the researchers capitalized upon the additional accuracy imparted by overlapping paired-end reads, accepting only those reads that matched perfectly in both directions. These added measures could explain why the frequency of detecting high-abundance false-positives was marginally lower than in our experimental setting, while recognizing that the barcode complexity of three was considerably less than 100. In a gene therapy context, such high-frequency false barcodes would misleadingly suggest the presence of non-existent clones. The detection of these false positives indicates that NGS analysis of complex barcode libraries has limited specificity. The mechanism by which these high-frequency systematic errors arise remains to be determined.

The potential for quantitative and unbiased analysis was an attractive advantage of barcoding over IS methodologies. For a barcode library of moderate complexity, containing 100 unique barcodes, only a crude approximation of clone size could be obtained. There was over one log-fold difference in the measured abundance of barcodes that had been mixed in equal proportions, despite minimizing PCR bias by using only 10 amplification cycles. Even barcodes that were present on the same plasmid molecule were detected at different abundances, to the extent that one barcode was readily detected, while the other was undetected or detected at levels below background, depending on the sequencing run. This indicates that the discrepancy in their relative abundance was independent of possible pipetting error during preparation of the plasmid mixture of 100 barcodes. Furthermore, analysis of barcodes that were mixed in equimolar proportions after amplification but before sequencing indicated that the discrepancy in the relative abundance of barcodes can arise solely during sequencing, although additional contributions of preferential amplification during PCR cannot be ruled out for other barcodes not tested in this way. Despite analyses of GC content of the 100 expected barcodes and predictions of their secondary structure, it remains unclear whether and how the barcode sequences themselves could contribute to discrepancies in their measured relative abundance.

Our investigation has shown that analysis of barcode libraries can give rise not only to false positives but also false negatives. In two independent Illumina sequencing runs, the same barcodes failed to be detected. Our empirical validation studies indicated that those barcodes were absent from both post-amplification samples derived from the 100-barcode plasmid library used for the first and second sequencing runs, even though these samples were generated in separate reactions using only 10 amplification cycles and the same primers homologous to constant sequences outside the barcode. The possibility that genuine clones could be missed during analysis of clinical gene therapy samples is of particular concern when monitoring for clonal dominance and malignancy. The danger that a malignant clone could be marked with a sequence that cannot be reliably detected limits the potential of this approach as a tool for clinical monitoring. Such a malignant clone may have to attain a greater level of dominance before expansion is identified as abnormal, risking further disease progression in a patient. Applications of barcoding that involve analyses of vector barcodes integrated in genomic DNA inescapably involve a PCR-based sample preparation stage in order to enrich for the barcode sequences and introduce the appropriate sequencing adaptors. For example, one or two rounds of amplification involving 25–45 PCR cycles have been used in other reports to retrieve barcode sequences from genomic DNA and introduce adaptor sequences (29–34). Even with a minimal number of PCR cycles, we observed biases caused by the failure of certain barcode sequences to amplify. Further optimizations of the PCR-based sample preparation step may ameliorate some of the biases we observed. For example, the performance of Kapa HiFi (Kapa Biosystems) is less affected by GC-rich and GC-low sequences compared to Phusion, and additives such as betaine and tetramethylammonium chloride may assist amplification of GC-rich and low-GC sequences (52,53). Again, depending on the particular application of barcoding, the failure to detect a minority of specific barcode sequences would not necessarily impugn the interpretability of the results.

The relatively higher error rates of NGS technologies have been identified as a challenge for applications of NGS and led to the development of error correction algorithms to facilitate data analysis. For example, sequence reads are known to be more accurate toward the beginning, and errors have been found to be associated with motifs such as G, inverted repeats and GGC (54–56). Many applications of NGS involve filtering sequence reads for known sequences or alignment to a reference sequence such as a genome.

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

These processes could account for the lack of precedence in the literature for systematic errors of the nature that we report. For example, around 60–90% of sequence reads typically pass alignment criteria during data analysis (25,56–59). It is likely that systematic errors are filtered out during these analytical steps.

Given the limitations to the analyzable degrees of complexity and clone size, and sensitivity and specificity, barcoding could not be used as a sole method for clinical monitoring of clonal diversity. Nonetheless, there could be some utility in barcoding if used in conjunction with conventional IS analysis, for providing a faster read-out, albeit crude. Furthermore, barcoding may still serve utility as an experimental tool for investigating hematopoiesis in small animal models, if library complexity fits within the limitations imposed by sequencing error. For example, a previous analysis of a similarly configured barcode identified 30–50 barcodes per transplanted mouse, which was sufficient to provide insights into hematopoietic differentiation (35). Other gene marking studies utilizing different barcode designs have detected fewer than 100 different barcodes in different lineages of transplanted mice (31,32). The reliability of such analyses could be further enhanced if fully characterized barcode libraries of moderate complexity, such as that in the present study, are used to validate data analysis strategies. It may also be possible to design fault-tolerant barcodes wherein the effect of sequencing error can be mitigated through the use of built-in features that enable error correction. Such a barcode could comprise, for example, tandem copies of known trinucleotide repeats. An error in 1 nt of the repeat could be corrected by reference to the other 2 nt. In practice, however, the construction of a complex barcode library with such a configuration may be difficult. Our findings imply that sequencing error may pose a challenge to the analysis of highly diverse TCR repertoires, which face similar difficulties, owing to the unknown number of TCR sequences of indefinite identity. For example, in a similar study to the present one, three single TCR sequences were analyzed using an Illumina platform and over 500 false TCR sequences were detected in each sample (51). Analytical strategies that take account of sequencing error are therefore required for identification of unique TCR sequences using NGS (60). Applications of TCR repertoire analysis that involve investigation of lower complexity TCR repertoires, such as monitoring for residual disease in patients treated for leukemia, especially when the malignant clonotype is known, are likely to be less affected by sequencing error (61,62).

In conclusion, we have shown that sequencing error limits the analysis of complex barcode libraries using contemporary NGS technology. Even with very high coverage of a defined library of moderate complexity, containing 100 barcodes, it was impossible to reliably distinguish all expected barcodes from false barcodes. Sequencing error thus imposes a limitation on the degree of complexity that can be resolved using NGS and highlights the importance of including known sequences in barcoding experiments. As has been demonstrated in other studies, applications where absolute distinction between true and false barcodes is not required because the acceptance of defined thresholds of error do not impugn the findings, analysis of barcoded samples can produce important biological insights (29,31,35). Furthermore, there may be applications for barcoding that involve lower orders of complexity and fall within the analyzable limit, such that the potential of contemporary NGS technology can be better capitalized upon for such applications. The use of paired-end sequencing in the configuration of a study such as the one described here would be expected to reduce background, and NGS technologies will continue to improve in terms of accuracy and reduction of systematic error. It is noteworthy that the resolution of complex barcode libraries constitutes a non-standard NGS application, and there will remain a need for vigilance of error pertaining to the use of methodologies lying outside their originally envisaged scope of application.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Margot Latham (The Children's Hospital at Westmead) for assistance in manuscript preparation.

FUNDING

Wenkart Foundation Grant; a National Health and Medical Research Council (NHMRC) Biomedical Postgraduate Scholarship [477112]; a Children's Medical Research Institute Scholarship to [C.T.D.]; an NHMRC project grant [1026710 to C.V.H. and I.E.A.]. Source of Open Access funding: Institution infrastructure funding.
Conflict of interest statement. None declared.

REFERENCES

1. Cavazzana-Calvo,M., Hacein-Bey,S., de Saint Basile,G., Gross,F., Yvon,E., Nusbaum,P., Selz,F., Hue,C., Certain,S., Casanova,J.L. *et al.* (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, 669–672.
2. Hacein-Bey-Abina,S., Hauer,J., Lim,A., Picard,C., Wang,G.P., Berry,C.C., Martinache,C., Rieux-Laucat,F., Latour,S., Belohradsky,B.H. *et al.* (2010) Efficacy of gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.*, **363**, 355–364.
3. Gaspar,H.B., Parsley,K.L., Howe,S., King,D., Gilmour,K.C., Sinclair,J., Brouns,G., Schmidt,M., Von Kalle,C., Barington,T. *et al.* (2004) Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet*, **364**, 2181–2187.
4. Gaspar,H.B., Cooray,S., Gilmour,K.C., Parsley,K.L., Adams,S., Howe,S.J., Al Ghonaium,A., Bayford,J., Brown,L., Davies,E.G. *et al.* (2011) Long-term persistence of a polyclonal T cell repertoire after gene therapy for X-linked severe combined immunodeficiency. *Sci. Transl. Med.*, **3**, 97ra79.
5. Aiuti,A., Slavin,S., Aker,M., Ficara,F., Deola,S., Mortellaro,A., Morecki,S., Andolfi,G., Tabucchi,A., Carlucci,F. *et al.* (2002) Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science*, **296**, 2410–2413.
6. Aiuti,A., Cattaneo,F., Galimberti,S., Benninghoff,U., Cassani,B., Callegaro,L., Scaramuzza,S., Andolfi,G., Mirolo,M., Brigida,I. *et al.* (2009) Gene therapy for immunodeficiency due to adenosine deaminase deficiency. *N. Engl. J. Med.*, **360**, 447–458.
7. Gaspar,H.B., Cooray,S., Gilmour,K.C., Parsley,K.L., Zhang,F., Adams,S., Bjorkegren,E., Bayford,J., Brown,L., Davies,E.G. *et al.* (2011) Hematopoietic stem cell gene therapy for adenosine deaminase-deficient severe combined immunodeficiency leads to long-term immunological recovery and metabolic correction. *Sci. Transl. Med.*, **3**, 97ra80.

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

8. Ott,M.G., Schmidt,M., Schwarzwaelder,K., Stein,S., Siler,U., Koehl,U., Glimm,H., Kühlcke,K., Schilz,A., Kunkel,H. *et al.* (2006) Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat. Med.*, **12**, 401–409.
9. Cartier,N., Hacein-Bey-Abina,S., Bartholomae,C.C., Veres,G., Schmidt,M., Kutschera,I., Vidaud,M., Abel,U., Dal-Cortivo,L., Caccavelli,L. *et al.* (2009) Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science*, **326**, 818–823.
10. Boztug,K., Schmidt,M., Schwarzer,A., Banerjee,P.P., Díez,I.A., Dewey,R.A., Böhm,M., Nowrouzi,A., Ball,C.R., Glimm,H. *et al.* (2010) Stem-cell gene therapy for the Wiskott-Aldrich syndrome. *N. Engl. J. Med.*, **363**, 1918–1927.
11. Aiuti,A., Biasco,L., Scaramuzza,S., Ferrua,F., Cicalese,M.P., Baricordi,C., Dionisio,F., Calabria,A., Giannelli,S., Castiello,M.C. *et al.* (2013) Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science*, **341**, 1233151.
12. Biffi,A., Montini,E., Lorioli,L., Cesani,M., Fumagalli,F., Plati,T., Baldoli,C., Martino,S., Calabria,A., Canale,S. *et al.* (2013) Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science*, **341**, 1233158.
13. Kalos,M., Levine,B.L., Porter,D.L., Katz,S., Grupp,S.a., Bagg,A. and June,C.H. (2011) T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci. Transl. Med.*, **3**, 95ra73.
14. Di Stasi,A., Tey,S.-K., Dotti,G., Fujita,Y., Kennedy-Nasser,A., Martinez,C., Straathof,K., Liu,E., Durett,A.G., Grilley,B. *et al.* (2011) Inducible apoptosis as a safety switch for adoptive cell therapy. *N. Engl. J. Med.*, **365**, 1673–1683.
15. Wang,G.P., Berry,C.C., Malani,N., Leboulch,P., Fischer,A., Hacein-Bey-Abina,S., Cavazzana-Calvo,M. and Bushman,F.D. (2010) Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood*, **115**, 4356–4366.
16. Hacein-Bey-Abina,S., Von Kalle,C., Schmidt,M., McCormack,M.P., Wulffraat,N., Leboulch,P., Lim,A., Osborne,C.S., Pawliuk,R., Morillon,E. *et al.* (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.
17. Hacein-Bey-Abina,S., Garrigue,A., Wang,G.P., Soulier,J., Lim,A., Morillon,E., Clappier,E., Caccavelli,L., Delabesse,E., Beldjord,K. *et al.* (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.*, **118**, 3132–3142.
18. Howe,S.J., Mansour,M.R., Schwarzwaelder,K., Bartholomae,C., Hubank,M., Kempski,H., Brugman,M.H., Pike-overzet,K., Chatters,S.J., Ridder,D. De *et al.* (2008) Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J. Clin. Invest.*, **118**, 3143–3150.
19. Stein,S., Ott,M.G., Schultze-Strasser,S., Jauch,A., Burwinkel,B., Kinner,A., Schmidt,M., Krämer,A., Schwäble,J., Glimm,H. *et al.* (2010) Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nat. Med.*, **16**, 198–204.
20. Gabriel,R., Eckenberg,R., Paruzynski,A., Bartholomae,C.C., Nowrouzi,A., Arens,A., Howe,S.J., Recchia,A., Cattoglio,C., Wang,W. *et al.* (2009) Comprehensive genomic access to vector integration in clinical gene therapy. *Nat. Med.*, **15**, 1431–1436.
21. Pule,M.A., Rousseau,A., Vera,J., Heslop,H.E., Brenner,M.K. and Vanin,E.F. (2008) Flanking-sequence exponential anchored-polymerase chain reaction amplification: a sensitive and highly specific method for detecting retroviral integrant-host-junction sequences. *Cytotherapy*, **10**, 526–539.
22. Paruzynski,A., Arens,A., Gabriel,R., Bartholomae,C.C., Scholz,S., Wang,W., Wolf,S., Glimm,H., Schmidt,M. and von Kalle,C. (2010) Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat. Protoc.*, **5**, 1379–1395.
23. Brady,T., Roth,S.L., Malani,N., Wang,G.P., Berry,C.C., Leboulch,P., Hacein-Bey-Abina,S., Cavazzana-Calvo,M., Papapetrou,E.P., Sadelain,M. *et al.* (2011) A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.*, **39**, e72.
24. Wu,C., Jares,A., Winkler,T., Xie,J., Metais,J.-Y. and Dunbar,C.E. (2013) High efficiency restriction enzyme-free linear amplification-mediated polymerase chain reaction approach for tracking lentiviral integration sites does not abrogate retrieval bias. *Hum. Gene Ther.*, **24**, 38–47.
25. Lam,H.Y.K., Clark,M.J., Chen,R., Chen,R., Natsoulis,G., O'Huallachain,M., Dewey,F.E., Habegger,L., Ashley,E.A., Gerstein,M.B. *et al.* (2012) Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.*, **30**, 78–82.
26. Ratan,A., Miller,W., Guillory,J., Stinson,J., Seshagiri,S. and Schuster,S.C. (2013) Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One*, **8**, e55089.
27. Suzuki,S., Ono,N., Furusawa,C., Ying,B.-W. and Yomo,T. (2011) Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS One*, **6**, e19534.
28. Benaglio,P. and Rivolta,C. (2010) Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region. *PLoS One*, **5**, e13071.
29. Nguyen,L.V., Makarem,M., Carles,A., Moksa,M., Kannan,N., Pandoh,P., Eirew,P., Osako,T., Kardel,M., Cheung,A.M.S. *et al.* (2014) Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. *Cell Stem Cell*, **14**, 253–263.
30. Cornils,K., Thielecke,L., Hüser,S., Forgber,M., Thomaschewski,M., Kleist,N., Hussein,K., Riecken,K., Volz,T., Gerdes,S. *et al.* (2014) Multiplexing clonality: combining RGB marking and genetic barcoding. *Nucleic Acids Res.*, **42**, e56.
31. Cheung,A.M.S., Nguyen,L. V, Carles,A., Beer,P., Miller,P.H., Knapp,D.J.H.F., Dhillon,K., Hirst,M. and Eaves,C.J. (2013) Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood*, **122**, 3129–3137.
32. Verovskaya,E., Broekhuis,M.J.C., Zwart,E., Ritsema,M., van Os,R., de Haan,G. and Bystrykh,L. V (2013) Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood*, **122**, 523–532.
33. Naik,S.H., Perié,L., Swart,E., Gerlach,C., van Rooij,N., de Boer,R.J. and Schumacher,T.N. (2013) Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*, **496**, 229–232.
34. Grosselin,J., Sii-Felice,K., Payen,E., Chretien,S., Roux,D.T.-L. and Leboulch,P. (2013) Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. *Stem Cells*, **31**, 2162–2171.
35. Lu,R., Neff,N.F., Quake,S.R. and Weissman,I.L. (2011) Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.*, **29**, 928–933.
36. Gerrits,A., Dykstra,B., Kalmykowa,O.J., Klauke,K., Verovskaya,E., Broekhuis,M.J.C., de Haan,G. and Bystrykh,L. V (2010) Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood*, **115**, 2610–2618.
37. Ginn,S.L., Liao,S.H.Y., Dane,A.P., Hu,M., Hyman,J., Finnie,J.W., Zheng,M., Cavazzana-Calvo,M., Alexander,S.I., Thrasher,A.J. *et al.* (2010) Lymphomagenesis in SCID-X1 mice following lentivirus-mediated phenotype correction independent of insertional mutagenesis and gammac overexpression. *Mol. Ther.*, **18**, 965–976.
38. Follenzi,A., Ailles,L.E., Bakovic,S., Geuna,M. and Naldini,L. (2000) Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences. *Nat. Genet.*, **25**, 217–222.
39. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.
40. Hall,T. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, **41**, 95–98.
41. Frey,B. and Suppmann,B. (1995) Demonstration of the Expand PCR System's greater fidelity and higher yields with a lacI-based PCR fidelity assay. *Biochemica*, **2**, 34–35.
42. McElroy,K., Thomas,T. and Luciani,F. (2014) Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb. Inform. Exp.*, **4**, 1.
43. Lou,D.I., Hussmann,J.A., McBee,R.M., Acevedo,A., Andino,R., Press,W.H. and Sawyer,S.L. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 19872–19877.

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016

44. Matochko,W.L. and Derda,R. (2013) Error analysis of deep sequencing of phage libraries: peptides censored in sequencing. *Comput. Math. Methods Med.*, **2013**, 491612.
45. Qu,W., Hashimoto,S.-I. and Morishita,S. (2009) Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Res.*, **19**, 1309–1315.
46. Zagordi,O., Klein,R., Däumer,M. and Beerenwinkel,N. (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.*, **38**, 7400–7409.
47. Kinde,I., Wu,J., Papadopoulos,N., Kinzler,K.W. and Vogelstein,B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 9530–9535.
48. Li,W., Fu,L., Niu,B., Wu,S. and Wooley,J. (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.*, **13**, 656–668.
49. Beerenwinkel,N., Günthard,H.F., Roth,V. and Metzner,K.J. (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.*, **3**, 329.
50. Minoche,A.E., Dohm,J.C. and Himmelbauer,H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.*, **12**, R112.
51. Nguyen,P., Ma,J., Pei,D., Obert,C., Cheng,C. and Geiger,T.L. (2011) Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics*, **12**, 106.
52. Quail,M.A., Otto,T.D., Gu,Y., Harris,S.R., Skelly,T.F., McQuillan,J.A., Swerdlow,H.P. and Oyola,S.O. (2012) Optimal enzymes for amplifying sequencing libraries. *Nat. Methods*, **9**, 10–11.
53. Ross,M.G., Russ,C., Costello,M., Hollinger,A., Lennon,N.J., Hegarty,R., Nusbaum,C. and Jaffe,D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, R51.
54. Kircher,M., Stenzel,U. and Kelso,J. (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.*, **10**, R83.
55. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
56. Nakamura,K., Oshima,T., Morimoto,T., Ikeda,S., Yoshikawa,H., Shiwa,Y., Ishikawa,S., Linak,M.C., Hirai,A., Takahashi,H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.
57. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
58. Degner,J.F., Marioni,J.C., Pai,A.A., Pickrell,J.K., Nkadori,E., Gilad,Y. and Pritchard,J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.
59. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
60. Britanova,O.V., Putintseva,E.V., Shugay,M., Merzlyak,E.M., Turchaninova,M.a., Staroverov,D.B., Bolotin,D.a., Lukyanov,S., Bogdanova,E.a., Mamedov,I.Z. *et al.* (2014) Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.*, **192**, 2689–2698.
61. Logan,A.C., Gao,H., Wang,C., Sahaf,B., Jones,C.D., Marshall,E.L., Buño,I., Armstrong,R., Fire,A.Z., Weinberg,K.I. *et al.* (2011) High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 21194–21199.
62. Wu,D., Sherwood,A., Fromm,J.R., Winter,S.S., Dunsmore,K.P., Loh,M.L., Greisman,H.a., Sabath,D.E., Wood,B.L. and Robins,H. (2012) High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.*, **4**, 134ra63.

Downloaded from http://nar.oxfordjournals.org/ at UNSW Library on May 18, 2016