Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing

Anders Ståhlberg[1,4], Paul M Krzyzanowski[2,4], Matthew Egyud[3], Stefan Filges[1], Lincoln Stein[2] & Tony E Godfrey[3]

[1]Department of Pathology and Genetics, Sahlgrenska Cancer Center, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden. [2]Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada. [3]Department of Surgery, Boston University School of Medicine, Boston, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.S. (anders.stahlberg@gu.se) or T.E.G. (godfreyt@bu.edu).

Published online 2 March 2017; doi:10.1038/nprot.2017.006

Detection of extremely rare variant alleles within a complex mixture of DNA molecules is becoming increasingly relevant in many areas of clinical and basic research, such as the detection of circulating tumor DNA in the plasma of cancer patients. Barcoding of DNA template molecules early in next-generation sequencing (NGS) library construction provides a way to identify and bioinformatically remove polymerase errors that otherwise make detection of these rare variants very difficult. Several barcoding strategies have been reported, but all require long and complex library preparation protocols. Simple, multiplexed, PCR-based barcoding of DNA for sensitive mutation detection using sequencing (SiMSen-seq) was developed to generate targeted barcoded libraries with minimal DNA input, flexible target selection and a very simple, short (~4 h) library construction protocol. The protocol comprises a three-cycle barcoding PCR step followed directly by adaptor PCR to generate the library and then bead purification before sequencing. Thus, SiMSen-seq allows detection of variant alleles at <0.1% frequency with easy customization of library content (from 1 to 40+ PCR amplicons) and a protocol that can be implemented in any molecular biology laboratory. Here, we provide a detailed protocol for assay development and describe software to process the barcoded sequence reads.

INTRODUCTION

The introduction of NGS has led to revolutionary capabilities in research and in clinical testing[1]. Sequencing of whole genomes, exomes or targeted genomic regions is now routine, and variant and/or mutant alleles can be identified with much higher sensitivity than with Sanger sequencing. Sensitivity of NGS depends on multiple factors, but detection of variants, in particular single-nucleotide variants, that occur at a frequency below ~1–3% remains challenging because of background noise[2]. In most applications, this level of sensitivity is adequate, but detection of very rare sequence variants is becoming increasingly important in many areas of research, including oncology, prenatal testing, infectious disease and forensics.

The main source of background in NGS is error introduced by DNA polymerases during library construction or sequencing. Approaches to overcome this error typically rely on deep sequencing and bioinformatics, barcoding of individual DNA template molecules or a combination of both[3–7]. Molecular barcoding of individual DNA template strands during NGS library construction results in the ability to track all sequencing reads back to a single original template. By aligning reads with the same barcode, it is then possible to differentiate between true variants and those resulting from *Taq* polymerase errors in any except the very first PCR cycle (**Supplementary Fig. 1**). One different and particularly elegant approach uses rolling circle amplification to generate concatenated copies of an original DNA strand such that sequence reads can again be traced back to the original DNA molecule[8]. These published approaches vary in their requirements and capabilities with regard to several important assay characteristics, such as DNA input and integrity requirements, flexibility of sequence targets, amount of the genome that can be sequenced and error correction capability. All require long and complex library preparation protocols. By contrast, SiMSen-seq incorporates barcodes using a novel PCR-based approach that facilitates low DNA input requirements, error correction to <0.1%, highly flexible target content and a very simple library construction protocol that is easy to implement in any molecular biology laboratory[9].

Development of the method

SiMSen-seq is made possible by two key innovations. First, the major problem associated with PCR-based barcoding is that the barcodes (typically random 8–14 mers) are not compatible with generation of specific PCR products. The random sequences virtually ensure nonspecific primer binding and formation of primer concatamers that can outcompete specific products. This problem is exacerbated when more than one amplicon is used in multiplex library construction. To overcome this, SiMSen-seq uses primers with built-in stem loop structures that hide or ‘protect’ the barcodes during the first PCR cycles. Subsequent PCR cycles are performed at higher temperatures to dissociate the stem and make the primers available for sequencing adaptors. Second, PCR components and conditions have been optimized using low primer concentrations in combination with extended annealing times, low enzyme concentrations, PCR additives and long PCR extension times to help improve specificity and produce more uniform amplicon yields when multiplexing. These conditions also provide another benefit in that the first PCR can be terminated, and the products can be used directly for adaptor PCR library construction without a cleanup step.

Here, we provide a detailed protocol for SiMSen-seq based on our previously published methods[9]. In brief, the protocol involves target primer design and testing, addition of barcoded hairpin primer sequences to the target primers and testing of the barcoded PCR primers, library preparation and library evaluation by fragment analysis, library sequencing and analysis of the sequence

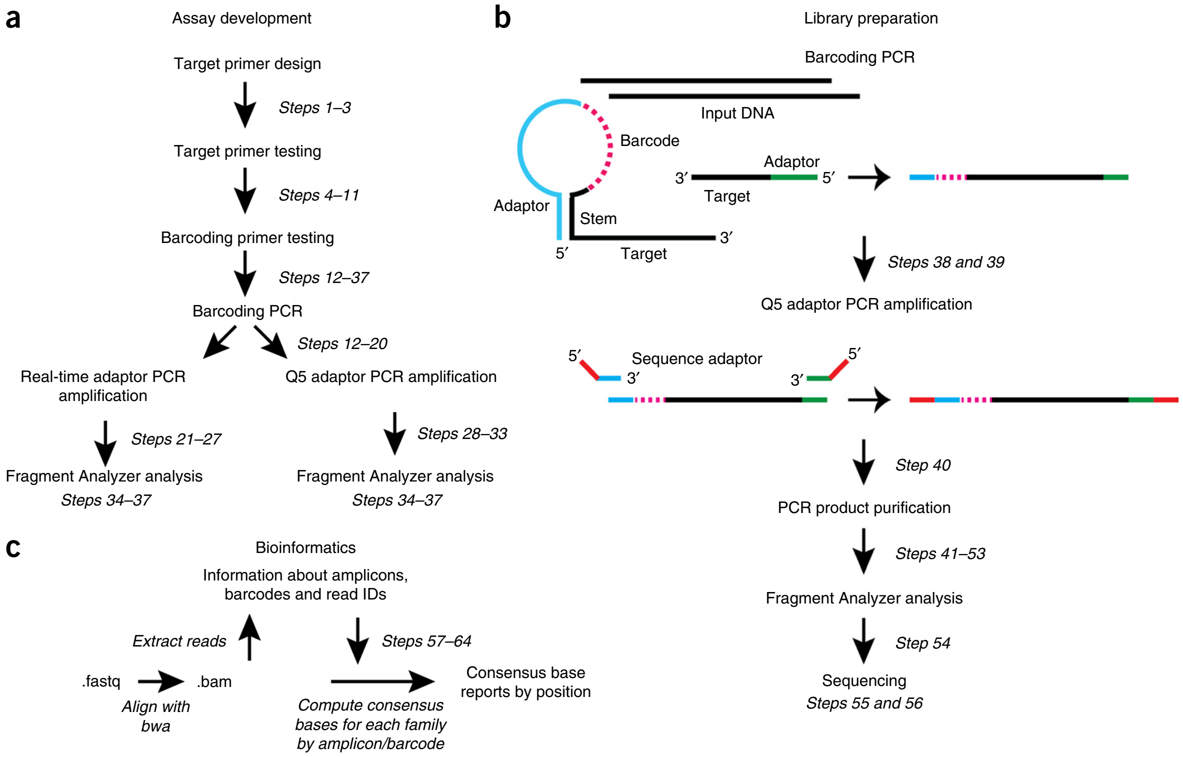

Figure 1 | Overview of the SiMSenSeq procedure. (**a**) Assay development (Steps 1–37). (**b**) Library preparation (Steps 38–56). (**c**) Bioinformatics using Debarcer (Steps 57–64). Several steps of assay development and library preparation are identical. Barcode primers and adaptor PCR primer sequences are illustrated with colors in b to illustrate the PCR amplicon composition. In the bioinformatics layout, generated data files/information are shown in regular type, whereas processes are shown in italic.

data to generate error-corrected reads (**Fig. 1**). The protocol also includes helpful information regarding anticipated results, data analysis and troubleshooting.

Overview of the method

PCR primer design and testing. SiMSenSeq starts with selection of the genomic regions of interest and design of PCR primers targeting these regions. For the most part, PCR primer design follows routine best practices and guidelines, with the goal being to obtain clean, specific PCR product with primers that function with high efficiency. Optimal annealing temperature for primer design is 60 ± 3 °C when using the Primer-BLAST tool described in the PROCEDURE (Steps 2 and 3). Another key consideration is the integrity of the target DNA source, as this affects the optimal PCR amplicon size. With high-quality DNA, optimal amplicon size is mostly defined by sequencing parameters such as read length and the choice of single- versus double-end reads. With lesser-quality DNA from sources such as formalin-fixed, paraffin-embedded tissues or from plasma, PCR amplicon size must be shortened accordingly to ensure that the majority of DNA strands are able to function as template, or it is possible that rare variants may not be represented in the final library. For plasma-derived DNA, we try to keep the amplicon size <80 bp, if possible.

Once designed, the target primers should be tested as described in PROCEDURE Steps 4–11, using quantitative real-time PCR followed by melting curve analysis using SYBR Green I detection chemistry. At this point, comparison with one of the control assays (TP1 and TP7) using primers listed in **Table 1**, run on the same plate, is a good guide, as cycle of quantification (Cq) values and plateau fluorescence for each DNA input should be very close to those of the control assay. Assays that show poor amplification characteristics or for which primers interact to produce a large amount of nonspecific PCR products in the no-template controls (NTCs) should be re-designed, if possible. We often find that moving a primer two or three bases 5′ or 3′ can make a marked difference in the results from NTCs, and ordering two or three slightly different forward and reverse primers and testing all combinations should be considered, as this may save time with little additional cost.

When a target primer design has been selected, the next step is to add the universal forward and reverse primer sequences that are used in the DNA barcoding PCR (**Table 1**). The universal hairpin sequence adds 63 bases in length to the forward primer, and the adaptor sequence adds 34 bases to the reverse primer. Thus, the resulting primers are typically 52–86 bases in length. Although our published work was performed with standard, desalted primers[9], recent data indicate that PAGE purification can substantially increase on-target yield and reduce nonspecific PCR products (data not shown). The magnitude of this effect seems to be amplicon-specific, however, with many standard primers performing very well. Thus, the cost and time of testing standard primers, and the possibility that PAGE purification will result in substantially improved on-target yield (and lower sequencing costs) for many assays, must be weighed against the additional cost of PAGE purification upfront.

The next step in individual primer testing is to make a test library using control DNA (Steps 12–37). This can be done using high-quality DNA, but we highly recommend testing a control DNA from a source similar to the intended sample. The first round of barcoding PCR is performed as in Steps 12–20 (typically

TABLE 1 | Primer sequences.

Name	Purpose	Primer sequence (5′–3′)
Barcode forward primer	Step 11	GGACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNNNNNNNNNNN**ATGGGAAAGAGTGTCC-forward target primer
Barcode reverse primer	Step 11	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-reverse target primer
Universal forward adaptor primer	Steps 21, 28	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Reverse index adaptor primer	Steps 21, 28	CAAGCAGAAGACGGCATACGAGAT**XXXXXX**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Target TP1 forward	Step 5	GTGGTGAGGCTCCCCTTT
Target TP1 reverse	Step 5	ACTGGGACGGAACAGCTTTG
Target TP7 forward	Step 5	CCTGGAGTCTTCCAGTGTGATG
Target TP7 reverse	Step 5	GACTGTACCACCATCCACTACAAC
Barcode TP1 forward	Step 11	GGACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNNNNNNNNNNN**ATGGGAAAGAGTGTCCGTGGTGAGGCTCCCCTTT
Barcode TP1 reverse	Step 11	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTGGGACGGAACAGCTTTG
Barcode TP7 forward	Step 11	GGACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNNNNNNNNNNN**ATGGGAAAGAGTGTCCCCTGGAGTCTTCCAGTGTGATG
Barcode TP7 reverse	Step 11	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACTGTACCACCATCCACTACAAC

Barcode and sample index barcode indexes are shown in bold font. X represents a given nucleotide according to the complete list of adaptor primers shown in **Supplementary Table 1**.

we run duplicate DNA and NTC reactions for each amplicon), and the products are used in a second-round quantitative PCR with Illumina sequencing adaptor primers—i.e., real-time adaptor PCR amplification testing (Steps 21–27). Cycling conditions are described in Step 26 and are designed to keep the hairpin stem open so that the adaptor primers can bind to the first-round PCR products. PCR is run for 40 cycles using SYBR Green I, and a melting curve is included at the end. Results are interpreted by comparison of amplification plots and melt curves from the DNA and NTC reactions (Step 27). If amplification plots and melting curves are conclusive (**Fig. 2**), it is safe to proceed to the final test with Q5 high-fidelity polymerase (below). If the results are not conclusive, the PCR products from the SYBR green PCR can be run on a Fragment Analyzer to determine how well an assay has performed. Finally, we typically repeat the adaptor PCR, this time with high-fidelity Q5 DNA polymerase (Q5 adaptor PCR amplification testing), and run the products on a Fragment Analyzer as a final test of the individual primer pairs. This is optional but recommended. Q5 adaptor PCR can be performed in parallel with the SYBR Green-I-based real-time PCR run, if desired.

After individual assays have been tested, it is also a good idea to evaluate any multiplex combinations that are required. This should be performed using an amount and type of DNA input similar to that which will be used for test samples. For this test, we do not recommend performing a real-time adaptor PCR step but instead proceeding directly to Q5 adaptor PCR after the barcode PCR, followed by fragment analysis. For low-level multiplexes (2–5 amplicons), fragment analysis can often provide a good idea of how well each amplicon is functioning (**Fig. 3**), but with higher-level multiplexes each amplicon may not be resolved and it is only possible to get a general impression of library content and quality. If it is critical to evaluate representation of all amplicons in a high-level multiplex, the best way is to generate the library and sequence it at low depth to determine the proportion of reads from each amplicon. We do not typically do this, as we find that most multiplexes provide adequate sequence data for all amplicons if they have passed individual testing.

SiMSen-seq library construction and sequencing. Sequencing libraries are generated using the same first-round barcoding PCR methods as described in the assay development phase. At this point, second-round adaptor PCR using SYBR Green I (Steps 21–27) is optional; one can proceed directly to Q5 adaptor PCR. SYBR green adaptor PCR may be helpful for two reasons: (i) comparison of amplification plots and melt curves with data from the assay development steps will give a good indication of whether the first-round PCR was successful and (ii) the amplification plot may be used to determine the number of PCR cycles to run in the actual library construction PCR using high-fidelity Q5 polymerase. For example, if PCR plateau is reached at *x* cycles in the SYBR green PCR, we would run *x*+2 cycles when using Q5 polymerase. This ensures good product yield while avoiding excessive PCR cycles that are not necessary and could result in additional nonspecific product formation. Alternatively, the SYBR Green I PCR step may be omitted, in which case the number of PCR cycles to run in the Q5 adaptor PCR step can be estimated from experience or by using the table in Step 32 as a guide. In either case, the actual library construction PCR is then performed with Q5 polymerase, as described in Step 40. Library PCR products are then cleaned up using AMPure XP beads (Steps 41–53) to remove nonspecific PCR products, and an aliquot is run on a Fragment Analyzer to evaluate and quantify the library (Step 54).

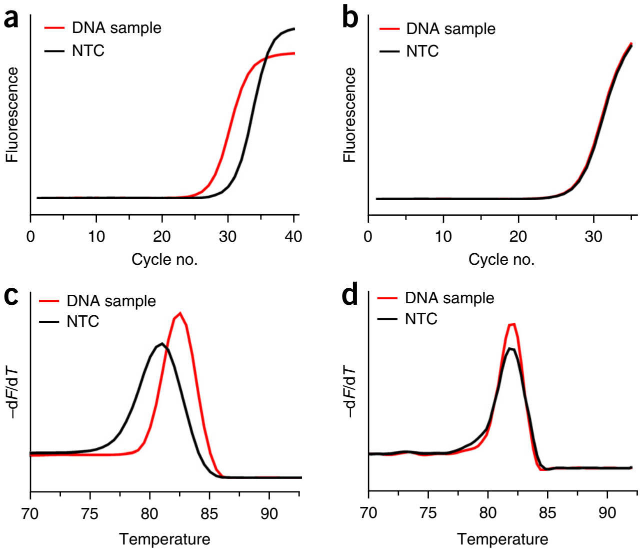

Figure 2 | Real-time adaptor PCR amplification testing. (**a**–**b**) Examples of amplification curves showing (**a**) a shift in Cq values between 20 ng of genomic DNA and an NTC and (**b**) an assay with no shift. (**c**,**d**) Example of melt curves for assays with (**c**) shifted and (**d**) overlapping melting temperatures between the DNA sample and the NTC. The following rules can be used to evaluate new assays: **a** + **c** = a good assay; **a** + **d** = probably a good assay (verify by Fragment Analyzer analysis); **b** + **c** = probably a poor assay, but this result is uncommon; **b** + **d** = probably a poor assay. d*F*/d*T*, the negative derivate of fluorescence versus temperature.

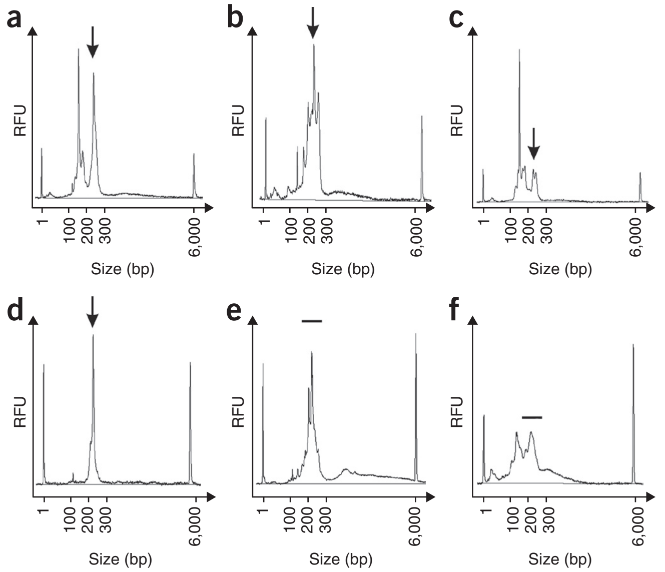

Figure 3 | Fragment analyzer examples for single- and multiplex libraries. (**a**–**d**) Single-plex libraries in which the arrows indicate the desired PCR product. (**e**,**f**) Multiplex libraries in which the horizontal bar indicates the expected size range of PCR products. All libraries were generated using 40–50 ng of high-quality human genomic DNA. **a**–**c** and **e**,**f** show unpurified libraries, whereas **d** shows a purified library. (**a**) Example of a good single-plex library with expected product size of 230 bp. Nonspecific PCR products are seen at ~160 bp but can easily be removed by purification (**d**). (**b**) Example of a poor single-plex library with expected product at 202 bp but with additional nonspecific products that cannot be removed by purification. Note that this library would still provide usable sequence data but with a lower on-target yield. (**c**) Example of a poor single-plex library with expected product size of 214 bp but with overall weak amplification and additional nonspecific products that cannot be removed by purification. It is unlikely that this library will provide good sequence data for detection of rare variants. (**d**) Purified version of library shown in **a**. Note that all low-molecular-weight PCR products have been removed, leaving a very pure library. (**e**) Example of a good three-plex library with expected product sizes of 210, 222 and 227 bp. Note that the 222- and 227-bp peaks cannot be resolved well, but the combined peak is higher than the 210-bp peak, indicating fairly even representation of the three products. (**f**) Example of an eight-plex library with expected product sizes ranging from 208 to 227 bp. Nonspecific products are present below 160 bp but can easily be removed by purification. Individual peaks cannot be resolved for this library. RFU, relative fluorescence units.

SiMSen-seq requires reading of libraries at high depth so that each unique barcode is read multiple times. Pooling of libraries for sequencing can be complex, as one must consider multiple factors, such as the desired consensus read depth, the number of haploid genomes present (DNA input), the estimated on-target read fraction for each library, the number of amplicons per library and the sequencer capacity (expected number of reads). For example, take a desired read depth of 20 per barcode (referred to as consensus 20) with a DNA input of 40 ng (~12,000 haploid human genomes). Assuming that all genome copies are uniquely tagged with one barcode each and that all reads are on-target, 240,000 reads would provide an average depth per barcode of 20. However, with three cycles of PCR barcoding, up to six unique barcodes can be incorporated into each genomic copy. One-third of the first-round PCR product is used in the adaptor PCR; therefore, on average two unique barcodes will be represented for each genome copy, for a total of 24,000 barcodes. In practice, however, not all the genomes will be represented; there will be uneven read depth per amplicon, not all reads will be on-target (this can be estimated from the fragment analysis run of the purified library) and if the average read depth is 20, many barcodes will be represented at <20× depth. Taking these points into consideration, a good starting point for this library would be a target of 480,000 reads per amplicon. With an on-target yield of anywhere >50%, this should provide adequate depth for consensus 20 mutation calling, with consensus 10 as a fallback option. Even at these very high read depths, multiple libraries can be pooled for sequencing. In this case, it is important to estimate the number of reads required per library using the example above as a guide. Using the number of total expected reads, one can then determine what fraction of the total reads is required for each library and pool the libraries accordingly. Sequencing can be performed on any Illumina sequencer. We currently recommend using single-end reads with an appropriate read length for your PCR product sizes.

Analysis of SiMSenSeq data

To facilitate the use of barcoded data generated by SiMSenSeq, we developed an analysis tool kit called Debarcer (de-barcoding and error correction). Debarcer is freely available for research use at https://www.github.com/oicr-gsi/debarcer.

Alignment and amplicon extraction.

Debarcer processes raw .fastq files containing SiMSen-seq barcoded adaptor regions using a combination of standard bioinformatic tools such as bwa, perl and R, as well as Bio-SamTools, to extract information from alignment files. The first step Debarcer takes is to identify the desired amplicons within raw, unaligned sequencing data. Untrimmed SiMSen-seq .fastq files are aligned using bwa to generate .bam files containing both aligned and unaligned reads. Output from the alignment step is used to define sequencing library specificity, which is the fraction of usable reads within the data obtained after sequencing. Generally speaking, library specificity is determined by the ability of SiMSen-seq primers to amplify the intended target regions and to minimize the amount of nonspecific products formed (as described in the assay development section of the PROCEDURE, Step 37). Alignment files generated by 'bwa mem' are parsed to identify any index-mapped reads that contain a properly oriented barcode and a SiMSen-seq adaptor region. These reads are categorized into families according to their barcodes and the start position of their alignment. Families are later assigned convenient aliases using an optional annotation table or automatically based on known gene symbols in the region—for example, a family associated with chr17:7577506 would be tracked as TP53_506. Debarcer populates its output directory with an alignment file and associated results for each SiMSen-seq .fastq file.

Some filtering also occurs at this stage. In our experience, we have found mutated barcodes that have arisen because of PCR errors in one of the library PCR cycles. Debarcer takes an extra step to identify these mutated barcodes and remove them before analysis. The most common evidence for a mutant barcode is pairs of barcodes with an edit distance of 1, for which the relative depth of both is highly dissimilar (i.e., 100:1). In these cases, the minor barcode probably arises from an error in replicating the major barcode. Although we find that on average 17–18% of all unique barcodes are mutated, the vast majority are found with less than three reads and cannot be used to compute consensus sequences. Although these barcodes are currently removed from the data set before analysis (they can be viewed in the Barcode Mask File), they represent ~1–3% of the data, and future versions of the pipeline may correct these errors to increase consensus sequence depths.

Consensus sequence building. Debarcer collects the read data for each amplicon and barcode (a 'sequence family'), and then, based on the alignment extracted from the .bam file, each base is indexed by genomic position. In addition, 'D' or 'I' is used to indicate deletions or insertions, respectively, detected at individual positions. Once all data for a given sequence family are indexed, a consensus sequence is generated. The default behavior is to assign each position the reference identity, unless a nonreference base is unanimously detected in a family with <20 total reads, or is detected at a level of ≥90% when the read depth is >20. These parameters can be modified by the user. As the number of final consensus sequences is related to the minimum raw read depth required to create a consensus sequence, more consensus sequences can be generated if a user decides to tolerate a lower number of raw sequences per barcode family. This decision comes at a cost of potentially allowing more false variants through to the final consensus reads. In our experience, data generated from families of depth 10 or higher (consensus 10) balance the need for PCR error correction and overall sequencing coverage.

Interpreting Debarcer reports. Debarcer can be used to evaluate the performance of SiMSen-seq assays and extract information about variants detected before and after error correction. The tool kit outputs several primary files that are referred to frequently: down-sampling figures to evaluate whether the sequence depth reached is adequate for a given SiMSen-seq assay; positional error bar plots to visualize raw and consensus error rates within a given genomic window; and position composition tables that contain more detailed allele counts at the raw and consensus levels across sequenced positions. These are organized in the top level of the output directory or tables/figures folders, according to their data type (**Supplementary Fig. 2**). Although several of these result files are discussed in further detail within the PROCEDURE section, Debarcer produces many other intermediate files that are useful for custom analyses. Further details beyond the scope of this article are described in the documentation that comes with the software.

Applications of SiMSen-seq

SiMSen-seq was developed initially for oncology research in which applications of rare mutation detection include analysis of tumor heterogeneity and identification of therapy-resistant clones[10]; detection of disease in biopsies, aspirates and cytology samples[11–14]; and early cancer diagnosis or recurrence monitoring using plasma, sputum, urine, stool or other bodily fluids[15–18]. Another area in which ultrasensitive mutation detection may be applicable is in prenatal testing. Detection and analysis of fetal DNA in maternal plasma have led to a revolution in noninvasive prenatal testing for Down syndrome and other disorders involving large chromosomal abnormalities[19]. Moving forward, detection of single-nucleotide variants specific to the fetus offers the potential to diagnose monogenic disorders early on in pregnancy without the risks associated with chorionic villus sampling or amniocentesis[20–22]. Other applications include scenarios in which DNA from a diseased tissue or pathogenic variant organism is present in a large background of normal DNA. Examples include detection of heteroplasmic mutations in mtDNA[23,24], detection of drug resistance mutations in viral diseases[25,26] and detection of donor DNA in the blood of transplant patients as an indication of organ rejection[27]. Finally, ultrasensitive variant detection will be useful in many fields, including forensics, toxicology, food testing evolution and others.

Comparison with other methods

Introduction of barcodes into NGS libraries has been reported previously, and methods include both ligation and incorporation of barcodes into PCR primers used for library generation[3–5,7,28–31]. Ligation approaches typically require larger amounts of DNA than does PCR, and also require several additional steps related to target capture for the genome regions of interest and associated cleanup and PCR steps. Furthermore, methods that include target capture are more applicable to multiple-kilobase or megabase target regions and are not typically used in scenarios in which library content must be customized from sample to sample. By contrast, PCR-based barcoding works with low DNA inputs and does not require target capture; in addition, library content is easily customizable. However, the Safe-SeqS protocol, as originally published, was performed with single amplicons, and the multiplexing capacity of this approach remains unclear. Furthermore, Safe-SeqS required multiple rounds of PCR, primer digestion and cleanup, a polyacrylamide gel purification and bead purification steps. In fact, with the exception of SiMSen-seq, all published approaches for NGS library barcoding use long, multistep workflows that are not only challenging to implement but also increase the risk of losing rare variant alleles as a result of inefficient ligation or capture, or during multiple cleanup steps.

Strengths and limitations of the method

One of the most important characteristics of SiMSen-seq, as compared with all other methods, is the simplicity of the NGS library construction protocol and the ease with which it can be implemented in any reasonably capable research laboratory. Once the assays are developed, SiMSen-seq library construction is simply composed of two rounds of PCR with a single cleanup step after the second PCR, and the entire library construction can be completed within 3–4 h.

Additional strengths of SiMSen-seq are that the first step is PCR-based, which facilitates the use of low DNA input (<5 ng), and that PCR amplicon sizes can be designed to work with both high- and low-integrity DNA (**Supplementary Figs. 3** and **4**). Both these characteristics can be useful when working with the limited amounts of fragmented DNA that are often obtained from plasma or serum. In this scenario, it is critical that the amplicon sizes be kept as short as possible to maximize the library yield and sensitivity[32]. Similarly, SiMSen-seq is very flexible, as the sequence target is defined by the PCR primer design and different amplicons can be multiplexed to generate unique NGS libraries as needed on an individual sample basis. The upper limit for multiplexing with SiMSen-seq is currently undetermined (we have tested up to 40 amplicons), but in theory high-order multiplexing should be possible as in standard NGS protocols such as the AmpliSeq product line from Thermo Fisher Scientific. However, design and testing of very large SiMSen-seq panels is time-consuming, and in the absence of commercially available panels it might not be the best approach for coverage of consistent, large target regions on many samples. Furthermore, barcoding strategies require very deep sequencing, and sequencing costs may become prohibitive as the size of the target DNA region increases. Thus, the ideal application of SiMSen-seq is probably one with relatively small target coverage on many samples or for which the target coverage is highly variable from sample to sample.

With regard to error correction, SiMSen-seq should provide data similar to those of other barcoding strategies that barcode a single strand of a DNA duplex, with background noise of <0.1% variant allele frequency at >99% of bases interrogated[9]. However, not all sequencing errors can be corrected by SiMSen-seq. For example, polymerase errors that occur in the barcoding PCR step are not corrected. This effect is minimized by the use of high-fidelity polymerases. One way to further reduce the impact of polymerase errors during PCR-based barcoding is to introduce barcodes into both DNA template strands as reported by others[7]. Although this is theoretically possible with the SiMSen-seq approach, we have not yet attempted it. Furthermore, some sequencing errors, such as chemically modified bases, are not polymerase-induced and cannot be corrected by barcoding in general.

In SiMSen-seq, we recommend using three PCR cycles to barcode template DNA—not the minimal number of two cycles. This is both a strength and a potential weakness. The strength is that using three cycles allows us to inactivate the barcoding PCR with a combined TE buffer dilution and protease treatment step instead of purifying the first-round PCR product. The use of three barcoding cycles results in up to eight barcoded molecules (with six unique barcodes) with forward and reverse adaptor sequences per original double-stranded template DNA molecule (**Supplementary Fig. 5**). One-third of the diluted PCR product is then used in the second-round PCR so that on average each original molecule should be represented by two barcoded molecules. This approach enables the rapid and simple workflow of SiMSen-seq. The potential weakness, however, is the possibility that a rare variant will be missed because of sampling error. If we assume that the number of molecules transferred from the first to the second PCR follows a hypergeometric distribution, the probability of transferring at least one of the eight possible barcoded molecules generated from any given template to the second PCR is ~96% with our recommended experimental setup. Theoretically, this slightly reduces the probability of detection of variant alleles if they are present at very low absolute copy numbers. Practically, however, the increased probability of a false-negative result from three barcode cycles and dilution versus two cycles and purification (with associated loss) is generally negligible.

Another consequence of omitting a purification step between the two rounds of PCR is the possibility that a 'new' barcode could be introduced during early cycles of the adaptor PCR. Thermodynamically, this is extremely unlikely, given the low concentration of the barcode primers carried over in the adaptor PCR (3.33 nM versus 400 nM for adaptor primers) and the annealing temperature difference used in the barcode PCR (62 °C) versus the adaptor PCR (72 °C). If a new barcode were introduced early enough in the second PCR to result in a consensus read, the impact would be to artificially alter the absolute variant allele fraction reported for each base in the amplicon. This effect is probably negligible, but if it is a concern then purification to remove the barcode primers following the barcode PCR would be an option.

Overall, when considering the multiple options for ultrasensitive sequencing, the key features of SiMSen-seq are low DNA input, ability to multiplex targets, flexible multiplexing, a <0.1% variant allele frequency background and a simple library construction protocol that requires no special equipment or expertise.

MATERIALS

REAGENTS

- Genomic DNA from the sample of interest. DNA from commercial providers, such as Human Genomic DNA (Roche Diagnostics, cat. no. 11691112001), may be used for assay development.
- AccuPrime *Taq* DNA polymerase, high fidelity, supplied with 10× buffer II (Thermo Fisher Scientific, cat. no. 12346086)
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A 63881) **! CAUTION** Beads contain sodium azide, which may form explosive compounds with heavy metals. Handle using appropriate safety equipment.
- BSA, 20 mg/ml, supplied in 10 mM Tris–HCl (pH 7.4), 100 mM KCl, 1 mM EDTA and 50% (vol/vol) glycerol (Thermo Fisher Scientific, cat. no. B14)
- Dilution Buffer E, 1×, 75 ml (Advanced Analytical, cat. no. DNF-495-0075)
- dNTP mix, 10 mM (each; Sigma-Aldrich, cat. no. D7295-20X.2ML)
- DPBS (Thermo Fisher Scientific, cat. no. 14190144)
- Ethanol 99.5% (vol/vol) (Kemetyl, cat. no. SN366915-06) **! CAUTION** Ethanol is flammable. Keep it away from ignition sources.
- Fragment Analyzer DNF-910-33-DNA-35-1500 bp kit (Advanced Analytical, cat. no. DNF-910-K0500) **! CAUTION** The

intercalating dye contains DMSO. Handle all reagents with appropriate safety equipment.
- Nuclease-free water (Thermo Fisher Scientific, cat. no. 15230147)
- Phusion Hot Start II high-fidelity DNA polymerase, supplied with 5× HF buffer (Thermo Fisher Scientific, cat. no. F-549L) ▲ CRITICAL Numerous high-fidelity DNA polymerases exist, but they vary highly in performance. We have had good experience with the AccuPrime *Taq* DNA polymerase and the Phusion Hot Start II high-fidelity DNA polymerase.
- Protease from *Streptomyces griseus*, type XIV, ≥3.5 U/mg, solid, powder (Sigma-Aldrich, cat. no. P5147-100MG) ! CAUTION It may cause serious eye, skin or respiratory irritation. Handle it with appropriate safety equipment.
- qPCR mix containing SYBR Green I—e.g., TATAA SYBR GrandMaster Mix (TATAA Biocenter, cat. no. TA01-625) or iQ SYBR Green Supermix (Bio-Rad, cat. no. 170888)
- Q5 Hot Start High-Fidelity 2× Master Mix (New England BioLabs, cat. no. M0494L)
- TE-buffer solution, pH 8.0 (Thermo Fisher Scientific, cat. no. AM9858)
- Oligonucleotides. See **Table 1** for a list of required oligos and their sequences. We order desalted and PAGE-purified primers at the lowest possible synthesis scale from IDT (http://www.idtdna.com/site).

EQUIPMENT
- Conical 15-ml tube (Thermo Fisher Scientific, cat. no. 339650)
- DynaMag-96 Side Magnet (Thermo Fisher Scientific, cat. no. 12331D)
- Fragment Analyzer (Advanced Analytical) ▲ CRITICAL The Fragment Analyzer outperforms comparable systems (e.g., Agilent Bioanalyzer 2100) in terms of peak resolution for SiMSen-seq libraries. This is critical to accurate quality control.
- Heating block (Eppendorf ThermoMixer C or equivalent)
- Illumina sequencing instrument (e.g., MiSeq or HiSeq 2000)
- Laboratory centrifuge (BioSan LMC-3000 or equivalent)
- MicroAmp Fast Optical 96-well reaction plate (Thermo Fisher Scientific, cat. no. 4346907)
- Microcentrifuge Safe-Lock tubes, polypropylene (Eppendorf, cat. no. 0030 120.086)
- Mini Microcentrifuge (Sigma-Aldrich, cat. no. CL S6766-1EA or equivalent)
- Pipettes and tips (Eppendorf or equivalent)
- Real-time PCR instrument (Thermo Fisher Scientific StepOnePlus Real-time PCR system or equivalent)
- Sealing tape, optically clear (Sarstedt, cat. no. 95.1994) ▲ CRITICAL Several plastic films exist. In contrast to the recommended film, most other films are difficult to remove, and some others adhere poorly and should be avoided.
- Thermal cycler (Bio-Rad T100 or equivalent)
- Vortexer (Scientific Industries VortexGenie 2 or equivalent)
- 96-well PCR plate, half skirt (Sarstedt, cat. no. 72.1979.202)
- Debarcer software. To download Debarcer and obtain up-to-date installation information, visit https://www.github.com/oicr-gsi/debarcer
- primer3 software (http://www.ncbi.nlm.nih.gov/tools/primer-blast/)

REAGENT SETUP

Protease preparation Dissolve 100 mg of *Streptomyces g.* protease powder in 5 ml of DPBS for a 667× solution. Store aliquots at −20 °C for up to 6 months. The 667× protease solution can be frozen–thawed at least 10 times. ▲ CRITICAL The protease solution should be prepared in advance.

Primer preparation Dissolve primers in TE buffer (pH 8.0) to reach a 100 µM stock concentration, gently mix and spin down at maximum speed in a minicentrifuge for 10 s at room temperature (18–23 °C). Forward and reverse barcode primers usually require an incubation step at 50–55 °C for 3–5 min using a heating block to completely dissolve. Subsequently vortex the sample, collect the solution at the bottom of the tube using centrifugation and store at −20 °C. Standard and adaptor primers are diluted to 10 µM each, whereas the barcode primer pool is prepared by mixing all primers of interest and diluting them to 1 µM each using nuclease-free water. Diluted primers stored at 4 °C should be used the same day (1 µM) or within a month (10 µM). Keep diluted primers in aliquots at −20 °C for up to 6 months. ▲ CRITICAL Dilute primer stock solutions and prepare primer pools in advance. Check all primer concentrations, especially barcode primers, after stock solutions are prepared. We sometimes find discrepancies between calculated and actual concentrations, and the primers are used at such low final concentrations that this is critical.

PROCEDURE

▲ CRITICAL To minimize the chance of contamination, prepare all reagents and master mixes in a PCR hood with UV sterilization. Load all template DNA material in a separate DNA loading area. Furthermore, all final PCR products and libraries should be handled in a post-PCR area.

▲ CRITICAL Use a thermal cycler with the lid heated to 105 °C for all PCR experiments in this protocol.

Assay development: target primer design ● TIMING 20 min per assay

1| Download target sequences for primer design. For example, use the UCSC Genome Browser (https://genome.ucsc.edu/). Select 'Genome Browser' in the task bar and then choose the appropriate genome version in the 'Genomes' tab. Both human GRCh37/hg19 and GRCh38/hg38 may be used. Enter the genomic position of interest and add 5 bp on both sides. Go to 'View' and select 'DNA' and add 200 bp at the 5′ and 3′ ends under 'Sequence Retrieval Region Options'. Retrieve the DNA sequence for the region by pressing 'get DNA'.

2| Design primers (Steps 2 and 3). Using primer3, paste the sequence of interest (from Step 1) into the 'PCR template' window. Fill in allowed 'Range' for forward and reverse primers to cover the sequence of interest. Use default settings, except as shown below.

Parameter	New setting
PCR product size (Primer Parameters)	60–80 bp (for degraded DNA)
Database (Primer Pair Specificity Checking Parameters)	Genome (reference assembly from selected organisms)
Max Self Complementarity (Advanced parameters)	3
Max Pair Complementarity (Advanced parameters)	3

▲ CRITICAL STEP For intact DNA, longer PCR assays can be used to cover larger sequences.

▲ CRITICAL STEP Several primer design tools exist with different features that can be applied in order to design primers with similar end results.

3| Retrieve primer suggestions by selecting 'Get primers'. Select and order the primer pair that displays the highest specificity and lowest self and self 3′ complementarity.
▲ **CRITICAL STEP**. If no primers are obtained, adjust the stringency of the primer design parameters. We avoid changing the primer melting temperature and PCR product length, instead choosing to adjust settings related to primer–dimer formation, primer specificity, GC content and primer length.
▲ **CRITICAL STEP** Some target sequences display a high degree of similarity, with unintended sequences that cannot be avoided by changing the primer design—e.g., in the case of pseudogenes. Amplification of these unintended sequence targets will consume sequencing capacity, but they can be deleted bioinformatically by sequence alignment.
▲ **CRITICAL STEP** We prefer to design and order several different primers targeting the same sequence for experimental testing. This can reduce assay development time by eliminating the synthesis and shipping delay associated with an iterative design, testing and ordering process.

Assay development: target primer testing ● TIMING **2 h**

4| Generate a dilution series with 3–5 genomic DNA concentrations ranging between 0.05 and 50 ng of genomic DNA/μl.

5| Assess target primer performance using quantitative real-time PCR on the genomic DNA dilution series with SYBR Green I detection chemistry. Analyze all samples, including the NTCs, as duplicates. At this point, we also recommend running assays for a reference assay for direct comparison with your new target primer designs. Target primers for two reference assays (TP1 and TP7) are shown in **Table 1**. Prepare the following master mix with all reagents except the DNA sample in a template-DNA-free area.

Reagent	Volume (μl)	Final
qPCR mix containing SYBR Green I (2×)	5	1×
Forward standard primer (10 μM)	0.4	400 nM
Reverse standard primer (10 μM)	0.4	400 nM
Nuclease-free water	2.2	–
Genomic DNA sample from Step 4, add in Step 7	2	1–100 ng
Total volume	10	–

6| Add 8 μl of master mix to a 96-well plate suitable for real-time PCR, and transfer the plate with the master mix to a template loading area.

7| Add 2 μl of genomic DNA sample (from Step 4) or 2 μl of water to each well to obtain a final volume of 10 μl.

8| Seal the plate with plastic film and spin down (1,200*g* for 10 s at room temperature—i.e., 18–23 °C) to collect the reaction mix at the bottom of each well.

9| Perform the PCR in a thermal cycler using the following temperature profile.

Cycle	Denature	Anneal	Extend	Melting curve
1	98 °C, 3 min	–	–	–
2–46	98 °C, 10 s	60 °C, 30 s	72 °C, 20 s	–
47	–	–	–	0.5 °C/5 s, from 60–95 °C

10| *Qualitative assessment of target primers*. Calculate the PCR efficiency of the standard curve using the following equation: PCR efficiency = 10^(1/–slope)

Determine the slope by plotting the cycle of quantification (Cq) value against the logarithmic (log10) starting DNA input. Assays that fulfill the following criteria are suitable to be tested as barcode primers: PCR efficiency >90%; negative water control samples do not generate any amplification curve within 34 cycles; only one PCR product with correct amplicon length is observed using the Fragment Analyzer (Steps 34–37).

▲ CRITICAL STEP Formation of correct PCR products can be evaluated by different methodologies, such as use of the Fragment Analyzer (Advanced Analytical), 2% (wt/vol) agarose gel electrophoresis or use of the Agilent 2100 Bioanalyzer system (Agilent Technologies).

? TROUBLESHOOTING

11| For those primer pairs that pass the quality control criteria, order barcode PCR primers ('barcode forward primer' and 'barcode reverse primer', see **Table 1**) and prepare stocks as described in the Reagent Setup. Barcode primers for two reference assays (TP1 and TP7) are shown in **Table 1**.

▲ CRITICAL STEP PAGE purification of barcode primers uniformly results in superior assays as compared with the use of desalted primers, but this difference is meaningful in only ~50% of assays. In ~15% of cases, PAGE purification provides a marked improvement in assay performance. We routinely use PAGE-purified primers for all assays.

Assay development: barcoding PCR ● TIMING 1.5 h

12| Prepare a barcode primer pool as described in the Reagent Setup; the number of primers will depend on the level of multiplexing in the desired library.

▲ CRITICAL STEP To test barcode primers, we recommend that each assay be tested individually, as well as in its final multiplex combination.

13| Prepare the barcoding reaction master mix with all reagents except the control DNA sample in a template- DNA-free area; if you are using AccuPrime *Taq* polymerase, high fidelity, use the first table, and if you are using Phusion Hot Start II high-fidelity DNA polymerase, use the second table. Note that volumes of water and DNA sample can be adjusted to obtain the final volume of 10 l. Analyze all samples, including the NTCs, in duplicate.

Reagent	Volume (µl)	Final
AccuPrime buffer II (10×)	1	1×
AccuPrime *Taq* polymerase, high fidelity (5 U/µl)	0.04	0.2 U
BSA (1 mg/ml)	0.5	0.05 mg/ml
Barcode primer pool (1 µM each primer)	0.4	40 nM
Nuclease-free water	Variable	–
DNA sample, add in Step 15	Variable	5–100 ng
Total volume	10	–

Reagent	Volume (µl)	Final
Phusion HF buffer (5×)	2	1×
Phusion Hot Start II polymerase (2 U/µl)	0.05	0.1 U
10 mM (each) dNTP mix	0.2	200 µM (each)
Barcode primer pool (1 µM each primer)	0.4	40 nM
Nuclease-free water	Variable	–
DNA sample, add in Step 15	Variable	5–100 ng
Total volume	10	–

▲ CRITICAL STEP According to the manufacturer and our experimental data, the Phusion DNA polymerase has a lower error rate than the AccuPrime DNA polymerase, and this may result in a further reduction in background noise in the sequencing data.

▲ CRITICAL STEP It is important to use hot start enzymes to reduce the formation of nonspecific PCR products during preparation.

▲ CRITICAL STEP For assay development, we recommend that 20–50 ng of control genomic DNA be used for the barcode PCR step. However, if it is known that <20 ng of DNA is available for your actual samples, we advise that the barcode PCR be performed with the corresponding amount of DNA.

14| Add master mix to a 96-well PCR plate and transfer the plate with the master mix to a template loading area.

15| Add DNA sample or water to each well to reach a final volume of 10 µl.

16| Seal the plate with plastic film and spin down (1,200*g* for 10 s at room temperature) to collect the reaction mix at the bottom of each well.

17| Perform the PCR in a thermal cycler using the following temperature profile. Note that at the beginning of cycle 5 protease will be added to inactivate the first PCR (Steps 18 and 19).

Cycle	Denaturation	Annealing	Elongation	Protease /inactivation	Hold
1	98 °C, 30 s	–	–	–	–
2–4 (3 cycles)	98 °C, 10 s	62 °C, 6 min	72 °C, 30 s	–	–
5	–	–	–	65 °C, 15 min	–
6	–	–	–	95 °C, 15 min	–
7	–	–	–	–	4 °C

▲ **CRITICAL STEP** Note that the annealing temperature for AccuPrime and Phusion DNA polymerases is 62 °C, as compared with the 60 °C annealing temperature used with regular DNA polymerases—i.e., Step 9. This reflects the manufacturer's instructions for annealing temperatures when using these enzymes.

18| In a 1.5-ml tube, prepare a working protease solution that will be used to inactivate the first PCR. Add 1.35 µl of 667× *Streptomyces g.* protease solution to 600 µl of TE-buffer (pH 8.0) and keep the solution on ice until required. This volume of protease solution is sufficient to inactivate 24 PCR samples.

19| At the beginning of cycle 5 in Step 17, open the thermal cycler lid and carefully remove the plastic film (while the 65 °C incubation step is running). Add 20 µl of working protease solution to each sample. Then seal the plate immediately with a new plastic film and close the thermal cycler. Keep the plate at 4 °C until it is used in adaptor PCR amplification (Step 24).

▲ **CRITICAL STEP** The protease treatment destroys the DNA polymerase, and this along with the dilution with TE effectively inactivates the barcoding PCR. By inactivating the PCR at 65 °C, the amount of nonspecific PCR products is minimized. If the PCR is inactivated at room temperature or on ice, the amount of nonspecific PCR product may increase to the extent that it affects the second adaptor PCR.

▲ **CRITICAL STEP** To avoid severe evaporation, the inactivation step should be performed rapidly, preferably within 3 min. We have not observed any problems caused by minor degrees of evaporation (<20%).

20| Check the assay performance by real-time PCR followed by melting curve analysis (Steps 21–27). Determine the amount of specific versus nonspecific PCR products on a Fragment Analyzer (Steps 28–37). Both tests can be performed in parallel, but we prefer to perform the real-time adaptor PCR amplification testing before deciding whether to continue with Q5 adaptor PCR amplification followed by Fragment Analyzer testing. To ensure optimal library quality, proceed with the barcoded DNA sample (Step 19) within the same day; overnight storage of barcoded DNA samples may decrease the amount of final library products.

Assay development: real-time adaptor PCR amplification testing ● TIMING 2.5 h

21| Prepare the following master mix with all reagents except the barcoded DNA sample in a template-DNA-free area. Adaptor primer sequences are shown in **Table 1** and **Supplementary Table 1**.

Reagent	Volume (µl)	Final concentration
qPCR mix containing SYBR Green I (2×)	5	1×
Universal forward adaptor primer (10 µM)	0.4	400 nM
Reverse index adaptor primer (10 µM)	0.4	400 nM
Barcoded DNA from Step 19, add in Step 23	4.0	–
Nuclease-free water	0.2	–
Total volume	10	–

▲ **CRITICAL STEP** For real-time adaptor PCR amplification testing, the same reverse index adaptor primer can be used for all samples, as samples are not forwarded to pooling and sequencing (Step 55).

22| Add 6 μl of master mix to a 96-well plate suitable for real-time PCR, and transfer the plate with the master mix to a template loading area.

23| Centrifuge (1,200*g* for 10 s at room temperature) the plate with the barcoded DNA sample to collect the reaction mix at the bottom of each well.

24| Add 4.0 μl of barcoded DNA sample (from Step 19) or barcoded NTC sample to each well to reach a final volume of 10 μl.

25| Seal the plate with plastic film and spin down (1,200*g* for 10 s at room temperature) to collect the reaction mix at the bottom of each well.

26| Perform the adaptor PCR in a real-time PCR instrument, using the following program.

Cycle	Denaturation	Annealing	Melting curve
1	98 °C, 3 min	–	–
2–41	98 °C, 10 s	80 °C, 1 s; 72 °C, 30 s; 76 °C, 30 s. All with ramping at 0.2 °C/s	–
42	–	–	0.5 °C/5 s, from 60 to 95 °C

27| Evaluate amplification plots and melting curves from the DNA and NTC reactions; ideally, one should see a two- to five-cycle difference in the Cq values for the 20–50 ng of genomic DNA input versus the NTCs, and a shift to a higher temperature in the melting curves for the DNA input reactions. Examples of good and poor results are shown in **Figure 2a–d**. The amplification curves can also be used to determine the number of cycles that are needed to generate a maximum amount of PCR products without running excessive cycles.
▲ CRITICAL STEP NTC reactions consistently produce nonspecific products with a fairly early Cq value. This is not a concern, as long as the DNA input curve is shifted two to five cycles to the left, as this is a strong indicator that a primer pair is working well. Similarly, a shift to the right in the melting curve is also a good indicator, but the absence of a shift, or a subtle shift, does not necessarily mean a failed assay, particularly if the amplification plots are shifted as described. If in doubt, the PCR products can be run on a Fragment Analyzer to determine how well an assay has performed (Steps 34–37).
? TROUBLESHOOTING

Assay development: Q5 adaptor PCR amplification testing ● TIMING 1.5 h

28| Prepare the adaptor PCR master mixes with all reagents except the barcoded DNA in a template-DNA-free area. Adaptor primer sequences are shown in **Table 1** and **Supplementary Table 1**.

Reagent	Volume (μl)	Final concentration
Q5 master mix (2×)	20	1×
Universal forward adaptor primer (10 μM)	1.6	400 nM
Reverse index adaptor primer (10 μM)	1.6	400 nM
Nuclease-free water	6.8	–
Barcoded DNA from Step 19, add in Step 31	10	–
Total volume	40	–

▲ CRITICAL STEP For Q5 adaptor PCR amplification testing, any reverse index adaptor primer can be used.

29| Add 30 μl of master mix to a 96-well plate suitable for PCR, and transfer the 96-well plate with the master mix to a template loading area.

30| Centrifuge (1,200*g* for 10 s at room temperature) the plate with the barcoded DNA sample to collect the reaction mix at the bottom of each well.

31| Add 10 µl of barcoded DNA sample (from Step 19) or NTC to each well to obtain a final volume of 40 µl. Seal the plate and spin down the samples to the bottom of each well (1,200*g* for 10 s at room temperature).

32| Determine the appropriate number of adaptor PCR cycles to perform, according to Step 27. Alternatively, adjust the number of cycles based on the level of multiplexing and the amount of input DNA, according to the following guide:

DNA (ng)	Multiplexing	No. of cycles
5	1	32–35
20	1	30–31
20	4	28–29
20	32	26–27
80	1	28–29

▲ **CRITICAL STEP** The number of cycles depends on the amount of input DNA. We have found 24–35 cycles of amplification to be sufficient in most cases. The reaction should reach the plateau phase to increase the total yield.

33| Perform the adaptor PCR amplification in a thermal cycler, using the following program.

Cycle	Denaturation	Annealing/elongation	Hold
1	98 °C, 3 min	–	–
2–*x*[a]	98 °C, 10 s	80 °C, 1 s; 72 °C, 30 s; 76 °C, 30 s. All with ramping at 0.2 °C/s	–
x+1	–	–	4 °C

[a] *x* is the number of cycles determined in Step 25.

■ **PAUSE POINT** The adaptor PCR product can be stored at −20 °C for several weeks.

Assay development: adaptor PCR library evaluation using the Fragment Analyzer ● TIMING 1 h per 12 samples

34| Prepare and load the fresh gel–dye mix, the inlet buffer plate and the marker plate according to manufacturer's instructions, using the DNF-910-33-DNA-35-1500 bp kit.

35| Add 2 µl of library sample (from Step 33) to 22 µl of Dilution Buffer E in a 96-well plate.
▲ **CRITICAL STEP** The input range of the DNA must be between 0.5 and 50 ng/µl. For higher concentrations, the sample must be diluted.

36| Seal the plate and vortex briefly. Spin down (1,200*g* for 10 s at room temperature) the samples to collect the reaction mix at the bottom of each well.

37| Run the Fragment Analyzer using the DNF-910-33-DNA-35-1500 bp kit.
▲ **CRITICAL STEP** If no clear, specific PCR product peak is observed, it indicates a poorly performing assay. Nonspecific PCR products are typically shorter, with a major peak ~170 bp.
▲ **CRITICAL STEP** We recommend that all individual assays be tested according to Steps 12–37. In addition, the final multiplex should be tested in the same way, except that the real-time adaptor PCR amplification testing (Steps 21–27) can be omitted.
? TROUBLESHOOTING

Library preparation: barcoding PCR ● TIMING 1.5 h

38| Repeat Steps 12–19 using the DNA of interest. We typically do not perform replicate assays at this point because of limited DNA availability. However, if the DNA is not limiting, the user may wish to run replicates and select the best library for sequencing based on fragment analysis results.

39| (Optional) Repeat Steps 21–26 to determine whether specific PCR products are formed, and identify the optimal number of adaptor PCR amplification cycles using real-time PCR.

Library preparation: Q5 adaptor PCR amplification ● TIMING 1.5 h

40| Repeat Steps 28–33 using the barcoded DNA from Step 38.
▲ CRITICAL STEP If you intend to pool multiple different libraries for sequencing (Step 55), which will most often be the case, it is important to use different reverse index adaptor primers for each library (see **Table 1** and **Supplementary Table 1**). These primers each contain a unique 6-base index sequence that allows sequence reads to be demultiplexed for downstream analysis of each individual library.

Library preparation: PCR product purification ● TIMING 30 min

▲ CRITICAL PCR product purification should be performed in a post-PCR area to minimize the chance of contamination in library generation.
41| Before beginning the purification, let the AMPure XP beads equilibrate to room temperature for at least 15 min, and then vortex the beads for 10 s. The entire PCR product purification procedure should be performed at room temperature.

42| Spin down (1,200*g* for 10 s at room temperature) the library samples from Step 40 to collect the reaction mix at the bottom of each well.

43| Add 38 µl of AMPure XP beads (1:1 ratio) to each sample and mix by carefully pipetting up and down until the solution is homogeneous.
▲ CRITICAL STEP It is important to maintain a bead/sample ratio of 1:1 (vol/vol) in order to purify products of the correct sizes without losing any final Q5 adaptor PCR product.

44| Incubate the samples for 5 min at room temperature to allow DNA binding to the magnetic beads.

45| Place the 96-well plate on a magnetic stand and incubate for 2 min or until the solution has cleared and all beads have been drawn to the magnet side of the wells.

46| While the plate remains on the magnetic stand, remove the supernatant without disturbing the bead pellet.
▲ CRITICAL STEP If beads are drawn up into the pipette, return the sample to the well and wait until the liquid clears to remove the supernatant again.

47| Wash the bead pellets with 200 µl of freshly prepared 70% (vol/vol) ethanol. Incubate at room temperature for 30 s and then aspirate the ethanol and discard.
▲ CRITICAL STEP It is important to perform this step with the reaction plate situated on the magnetic stand. Do not disturb the separated magnetic beads.

48| Repeat the washing (Step 47) with a fresh 200 µl of ethanol.

49| If needed, remove residual ethanol using a P10 pipette.
▲ CRITICAL STEP Be sure to remove all the ethanol from the bottom of the well, as residual ethanol may affect downstream analysis. If the beads are dried for too long, they will be difficult to resuspend, which may cause a reduced yield.

50| Take the plate off the magnetic stand. Add 20 µl of TE-buffer (pH 8.0) to each well and resuspend the beads by pipetting up and down 10 times.

51| Incubate the plate at room temperature on the bench for 2 min.

52| Place the plate back onto the magnetic stand and incubate for another 2 min.

53| Transfer the supernatant to a new 96-well plate.

Library preparation: library quality control using the Fragment Analyzer ● TIMING 1 h

54| Repeat Steps 34–37 using the purified library from Step 53.
▲ CRITICAL STEP A good library should contain only products of the known amplicon sizes after purification.
? TROUBLESHOOTING

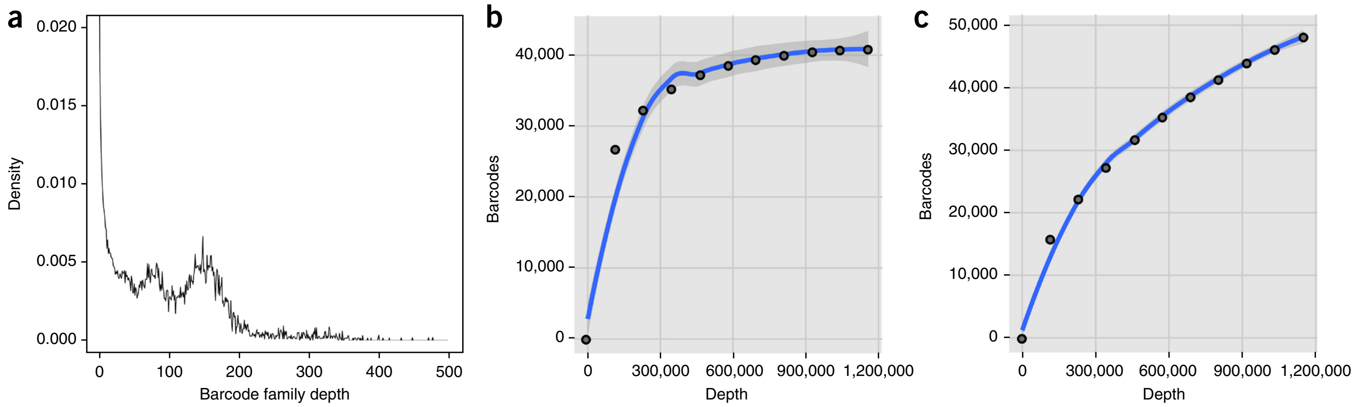

Figure 4 | Use of Debarcer to inform raw and consensus sequencing depth. (**a**) Distribution of barcode family depths found in one SiMSen-seq library. (**b**) Down-sampling plot showing saturation of detected barcodes, indicating that an appropriate sequencing depth has been reached. (**c**) Down-sampling plot indicating that additional barcodes may be detected with additional sequencing depth.

Sequencing ● TIMING 1–2 d

55| Pool libraries from Step 53 for Illumina sequencing, according to the manufacturers' instructions.
▲ **CRITICAL STEP** It is important to consider the amount of starting DNA, number of amplicons and estimated on-target yield for each library when pooling. These issues are discussed in the 'Overview of the method' section, 'SiMSen-seq library construction and sequencing' section, paragraph 2.

56| Sequence the pooled libraries on any Illumina sequencer according to the manufacturers' instructions. We currently recommend using single-end reads with an appropriate read length for your anticipated PCR product sizes.

Data analysis using Debarcer ● TIMING 0.5–6 h

57| Download and test Debarcer. Debarcer is available at GitHub (https://www.github.com/oicr-gsi/debarcer) and is bundled with a small test data set that can be used to verify that Debarcer functions on your system. This test data set should run to completion in under 5 min on any modern system.

58| (Optional) Configure `debarcer.conf` in the top level of your output directory. Debarcer checks for this file when starting and reads in assay-specific parameters—for example, a config file containing

```
plexity=8
```

instructs Debarcer to output the top eight detected amplicons, according to abundance.

59| Run Debarcer on your SiMSen-seq data. The minimal arguments required for Debarcer to process a SiMSen-seq. fastq file are the path to the .fastq file itself ('-f'), an output directory ('-o') and a run flag ('-r'), as follows:

```
runDebarcer.sh -r -f [/directory/filename.fastq] -o [.]
```

Debarcer itself requires no further user interaction and produces output files organized by type. Currently, this consists of tables and figures, as shown in **Supplementary Figure 2**. At a minimum, there are three main parameters that should be considered while evaluating a SiMSen-seq assay: the on-target yield of sequenced amplicons; the quantity of consensus sequences produced from a given data set; and the target depth of a sequencing run.

60| *Determine amplicon yield*. 'Amplicon Yield (Amplicon Reads/Raw reads)' indicates how much of the total data are analyzable and is generally related to the level of on-target PCR products in the reaction. Information about this parameter is found in the 'Summary Statistics file', which reports this value directly for the entire library.

61| *Optimization of the consensus depth to use as a cutoff*. The 'Consensus Sequence Yield' refers to the number of consensus sequences identified using a given consensus depth (consensus 10, consensus 20 and so on). The consensus sequence yield determines the lower limit of sensitivity of your assay and is primarily determined by the minimum number of raw reads used to compute a consensus sequence.

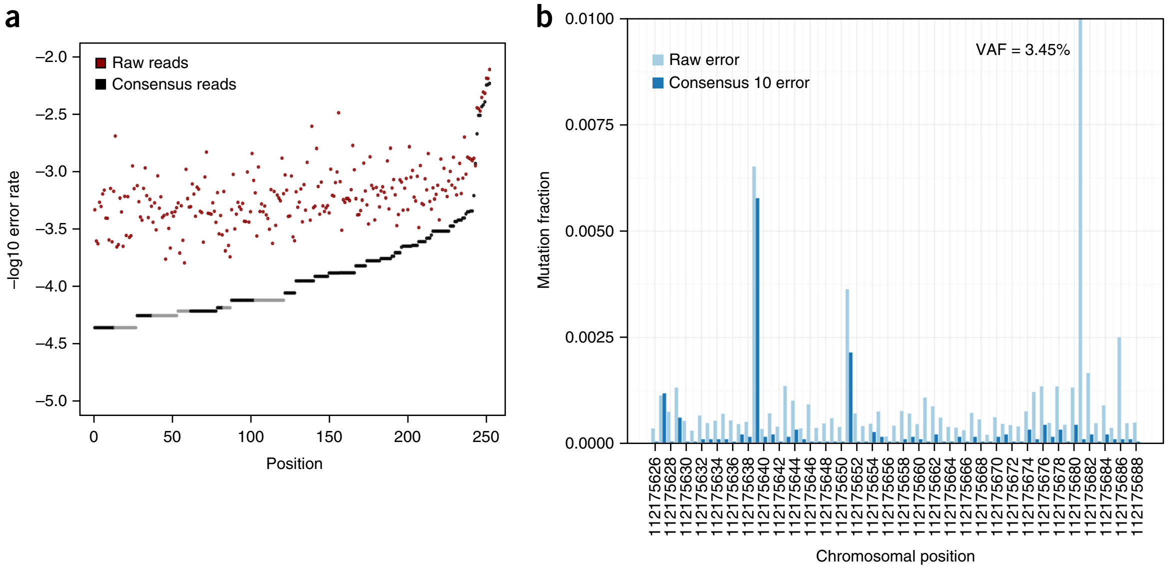

Figure 5 | Visualizing SiMSen-seq error correction with Debarcer. (**a**) Scatterplot showing the raw and corrected error rates for all positions in a five-plex library. Genomic positions are sorted by ascending consensus error rate. (**b**) A mutation bar plot for a single APC amplicon, showing raw and consensus 10 data. In the raw data, three potential low-level variants can be observed, with the highest variant allele frequency (VAF) being ~3.5%. In the corrected consensus 10 data, this variant is eliminated, leaving two possibly mutated positions, which are both well above the background error level.

Debarcer produces a 'Consensus Statistics' file that reports the count of consensus sequences for all amplicons across a range of consensus depths, as well as a 'Depths per Barcode' plot to visualize the range of practical consensus depth cutoffs (**Fig. 4a**). Using barcode families with high consensus depths results in greater ability to eliminate *in vitro*-generated base substitutions, at the expense of consensus read depth.

62| *Determination of sequence saturation*. The final critical output to check is whether the appropriate sequencing depth has been reached for a given SiMSen-seq assay (whether single-plex or multiplex). Debarcer produces 'Down-sampling plots', which chart the number of consensus sequences detected using different subsamplings of the alignment file (**Fig. 4b**,**c**).

Ideally, a horizontal plateau should be apparent in the down-sampling plot, indicating that the chosen library sequencing depth is sufficient to identify most, if not all, possible consensus sequences (**Fig. 4b**). An extensive plateau region suggests that the sample is being unnecessarily oversequenced.

By contrast, if the down-sampling plot ends with an upward slope (**Fig. 4c**), it suggests that additional sequence coverage of the same sequencing library would yield a greater number of consensus sequences, thereby increasing the limit of detection for variant alleles in the collapsed data.

63| (Optional) Verify overall correction performance of the library using the 'Amplicon Error' plots (**Fig. 5a**). These dot plots can be used to quickly decide two important features of a data set. First, most positions should show consensus read error rates lower than the raw error rate, and it is important to understand that polymerases do not replicate all positions and sequence contexts with the same fidelity; thus, there will be different correction factors for individual bases across your data set.

Occasionally, 'Amplicon Error' plots will show consensus error rates that are actually higher than the raw error rates; this is generally the result of low DNA input or insufficient sequence coverage, both of which lead to sequences being collapsed to a small number of barcode families. Both these effects can also be identified at Steps 61 and 62.

The second feature viewed through the 'Amplicon Error' plots is the lowest consensus sequence error rate for each amplicon. These points are plotted either in black, indicating that an error was found; or gray, indicating that no consensus sequence errors were found for that position. In these cases, the number of consensus sequences computed determines the lowest reportable error rate, and having many of these positions in a data set suggests similar problems with input DNA or sequence coverage, as described above.

64| *Viewing of data for specific genomic positions*. Debarcer outputs several files from which information on individual variants can be drawn directly. In addition to the 'Barcode Depth' and 'Downsampling' figures described in Steps 61 and 62, respectively, three main sources of information that users would be interested in are the 'Positional Error Bar' plots, the 'Position Composition' tables and the 'Amplicon Error' plots. As these data are used to compare raw sequence data with corrected data, and the optimal consensus depth for each SiMSenSeq assay will vary, Debarcer generates data sets for a range of consensus sequence depths (3, 10, 20, 30) by default.

The most common expected use for 'Positional Error' bar plots is to quickly identify whether a variant at a position of interest is a true variant or whether it represents a noisy genomic position. An example of this is shown in **Figure 5b**. The raw data for **Figure 5** would be found in the corresponding 'Positional Composition' table.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 2**.

TABLE 2 | Troubleshooting table.

Step	Problem	Possible reason	Solution
10	Low PCR efficiency	Poor primer binding	Re-design primers
		Presence of PCR inhibitor	Prepare good-quality DNA
		Poor dilution curve	Remake the dilution series. Reproducible DNA dilutions can be performed by diluting the DNA in 1 mg/ml BSA
	Formation of primer–dimers	Primers with complementary sequences	Re-design primers in Primer3, decrease maximum self/pair complementarity in the parameter settings
	Formation of nonspecific PCR products	Primers amplify unintended targets	Re-design primers, and increase target stringency settings
	Specific PCR products found in negative controls	Contamination	Clean all working areas, and change all possible solutions. All negative controls must be clean from template DNA. Make sure to have physically separated workplaces for preparing template-DNA-free solutions, template loading area and post-PCR working area
27	No shift in Cq values between positive and negative samples	Too much nonspecific PCR product formed	Re-design the primers
			If good primers are hard to obtain, check PCR products on the Fragment Analyzer. Even if the designed primers are not perfect, they may be specific enough for your application (test whether your intended DNA concentration results in correct PCR products)
		Degraded DNA/presence of PCR inhibitors	Use good-quality DNA. Make sure to use 20–50 ng of DNA to see a shift
	No shift in melting temperature between positive and negative controls	No specific PCR products are formed	Re-design the primers
		The melting temperatures of some specific PCR products are identical to the melting temperature of the nonspecific PCR product	Check PCR products using Fragment Analyzer
			Use more DNA; if more DNA is used you will observe a clear shift in Cq values between positive and negative samples, indicating that specific PCR products are actually formed, regardless of the melting curves
	Specific PCR products found in negative controls	Contamination	Clean all working areas, and change all possible solutions. All negative controls need to be free of template DNA. Make sure to have physically separated work places for preparing template-DNA-free solutions, template loading area and post-PCR working area

(continued)

TABLE 2 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
37	Weak/no amplification of specific product	Incorrect DNA input or DNA quality is poor	Check DNA concentration and integrity
		Primers functioning poorly	Although unusual, addition of adaptor sequences can cause good target primers to work inefficiently. Try using PAGE-purified primers or redesign target primers
		Barcode or adaptor primer stocks degraded	Follow instructions for primer handling in the Regent Setup. Repeat assay with new primer dilutions
	Nonspecific products close to desired products	Interaction between target and adaptor sequences or between primers in a multiplex	Adjust the bead ratio to remove as much nonspecific product as possible during purification. If performing a multiplex assay, consider leaving out each primer pair sequentially to identify the "toxic" primers and redesign or omit. If specific product is clearly present, consider sequencing, but with increased library representation to overcome low on-target yield
54	Little or no product after purification	Beads were allowed to dry out too much or beads were accidentally pipetted out during wash steps	Repeat purification, being careful to follow all instructions in Steps 41–53 and those supplied with the beads

● TIMING

The assay development protocol takes ~1 d of hands-on time to complete. However, two rounds of primers need to be purchased and delivered before completely optimized assays are ready. The entire process therefore typically takes 1.5–2 weeks, as turnaround time for PAGE-purified primers can be a week or more. Redesigns of failed target primers add an extra 2–3 d for 20–30% of assays but can often be avoided by designing and ordering multiple variations of target primers the first time around.

Target primer design and testing
Steps 1–3, target primer design: 20 min
Steps 4–11, target primer testing: 2 h
Barcode primer testing (Steps 12–37)
Steps 12–20, barcoding PCR: 1.5 h
Steps 21–27, real-time adaptor PCR amplification testing: 2.5 h
Steps 28–33, Q5 adaptor PCR amplification testing: 1.5 h
Steps 34–37, adaptor PCR library evaluation using the Fragment Analyzer: 1 h
Library preparation (Steps 28–54; can be completed within half a day)
Steps 38 and 39, barcoding PCR: 1.5 h
Step 40, Q5 adaptor PCR amplification: 1.5 h
Steps 41–53, PCR product purification: 30 min
Step 54, library quality control using the Fragment Analyzer: 1 h
Sequencing and data analysis
Steps 55 and 56, sequencing: 1–2 d.
Steps 57–64, data analysis: usually 0.5–6 h, depending on the amount of sequence data generated in any given run

ANTICIPATED RESULTS

Target primer performance After target primer design and testing (Steps 1–10), we anticipate that ~70% of the designed and tested primers will display high PCR efficiency (>90%) and form no/minor amount of off-target PCR products.

Characteristics of SiMSen-seq libraries (Steps 37 and 54) Unpurified single-plex libraries should show a strong product of expected size (±10%) with little or no additional peaks close by using the Fragment Analyzer (**Fig. 3a**). Strong peaks in the 160- to 170-bp range are expected, and these represent primer–dimer products that are subsequently removed by bead purification (**Fig. 3d**). Note that a total of 150 bases are added to the target amplicon size during the two rounds of PCR, and thus even very small target amplicons still result in library products that can easily be distinguished from the 160- to 170-bp primer–dimer products. Poorly performing assays typically have multiple additional product peaks close to those of the desired product and/or show poor amplification of the desired product (**Fig. 3b,c**). These assays may still provide adequate sequence data, but on-target yield will be lower and the library pooling should take this into account to ensure adequate representation.

Multiplex libraries should have characteristics similar to those of single-plex libraries: strong peaks at the desired sizes and no additional peaks other than primer dimers. For two- to five-plex libraries, the individual products may be distinguishable on the fragment analysis (**Fig. 3e**), but for more complex libraries this may not be the case and instead a broad 'peak' encompassing the expected product size range is typical (**Fig. 3f**). Minor peaks or broader than expected product peaks are also common (**Fig. 3e**) and may be a result of interaction between products or a secondary structure caused by the hairpin adaptor sequences. These are not a concern.

Purification of the final library is expected to result in an electropherogram with the desired PCR products and minimal amounts of primer–dimer or other nonspecific PCR products (**Fig. 3d**).

Characteristics of sequence data (Steps 60–64) Three key characteristics of the sequence data are expected of a successful library construction and sequencing run. First, the fraction of on-target reads for any library is typically >50%, and for single-plex reactions using PAGE-purified barcode primers we expect >85% on-target reads for most assays. Although not critical to the generation of successful consensus reads and error correction, higher on-target read fractions are desirable, as they result in reduced sequencing costs and generally indicate that all aspects of library construction functioned well. Second, the number of unique barcodes represented in the consensus reads at any given consensus depth is a critical indicator of good barcode incorporation and defines the lower limit of detection sensitivity. Ideally, an average of two barcodes should be represented for each original double-stranded DNA template molecule in the reaction (e.g., ~600 barcodes per ng of human DNA). In practice, issues such as template availability, PCR efficiency and sampling variability result in <100% of the expected barcode representation. In our experience, good libraries result in 300–600 unique barcodes per ng of human DNA when read at the desired consensus depth. When evaluating this, however, it is critical to also take into account the raw sequence depth to ensure that all barcodes are being read with adequate depth (see the 'Optimizing sequencing depth' section). Finally, consensus read sequencing error should be substantially lower than the raw read sequencing error at almost all bases. The error correction factor varies by base position and is driven to a large extent by the raw read error and the DNA input amount, but the average correction should be ~7- to 8-fold for a 50-ng DNA input. In addition, 40–60% of bases are expected to show zero consensus read error, and ~99% of base positions should have a <0.1% consensus read error. This is demonstrated in **Figure 5b**, in which multiple positions demonstrate raw read sequencing error >0.1%, including four bases >0.25%, and one at ~3.5%. Subsequent error correction, although variable by base position, yields a consensus read sequencing error that is several-fold lower at most base positions, and that in some cases is effectively 0%. This correction leaves two prominent base positions well above the consensus background error, indicating possible mutations.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS This work was supported by the National Institutes of Health (R21CA172999 to T.E.G.), the Swedish Cancer Society (to A.S.), the Swedish Childhood Cancer Foundation (to A.S.), the Sahlgrenska Academy at the University of Gothenburg (to A.S.).

AUTHOR CONTRIBUTIONS A.S. and T.E.G. conceived and designed the protocol; P.M.K. and L.S. developed the Debarcer software for data analysis; and M.E. and S.F. contributed to protocol development. A.S., P.M.K., M.E., S.F. and T.E.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare competing financial interests: details are available in the online version of the paper.

1. ten Bosch, J.R. & Grody, W.W. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J. Mol. Diagn.* **10**, 484–492 (2008).
2. Fox, E.J., Reid-Bayliss, K.S., Emond, M.J. & Loeb, L.A. Accuracy of next generation sequencing platforms. *Next Gener. Seq. Appl.* **1** (2014).
3. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
4. Schmitt, M.W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 14508–14513 (2012).
5. Flaherty, P. *et al.* Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.* **40**, e2 (2012).
6. Newman, A.M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
7. Kennedy, S.R. *et al.* Detecting ultralow-frequency mutations by duplex sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
8. Lou, D.I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 19872–19877 (2013).
9. Stahlberg, A. *et al.* Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Res.* **44**, e105 (2016).
10. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
11. Wu, H.H. *et al.* Utilization of cell-transferred cytologic smears in detection of EGFR and KRAS mutation on adenocarcinoma of lung. *Mod. Pathol.* **27**, 930–935 (2013).
12. Boadas, J. *et al.* Clinical usefulness of K-ras gene mutation detection and cytology in pancreatic juice in the diagnosis and screening of pancreatic cancer. *Eur. J. Gastroenterol. Hepatol.* **13**, 1153–1159 (2001).
13. Ohori, N.P. *et al.* BRAF mutation detection in indeterminate thyroid cytology specimens: underlying cytologic, molecular, and pathologic characteristics of papillary thyroid carcinoma. *Cancer Cytopathol.* **121**, 197–205 (2013).

14. Malapelle, U. *et al.* EGFR mutations detected on cytology samples by a centralized laboratory reliably predict response to gefitinib in non-small cell lung carcinoma patients. *Cancer Cytopathol.* **121**, 552–560 (2013).
15. Hoque, M.O. *et al.* High-throughput molecular analysis of urine sediment for the detection of bladder cancer by high-density single-nucleotide polymorphism array. *Cancer Res.* **63**, 5723–5726 (2003).
16. Thunnissen, F.B. Sputum examination for early detection of lung cancer. *J. Clin. Pathol.* **56**, 805–810 (2003).
17. Diehl, F. *et al.* Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. *Gastroenterology* **135**, 489–498 (2008).
18. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra68 (2012).
19. Lo, Y.M. & Chiu, R.W. Genomic analysis of fetal nucleic acids in maternal blood. *Annu. Rev. Genomics Hum. Genet.* **13**, 285–306 (2012).
20. New, M.I. *et al.* Noninvasive prenatal diagnosis of congenital adrenal hyperplasia using cell-free fetal DNA in maternal plasma. *J. Clin. Endocrinol. Metab.* **99**, E1022–E1030 (2014).
21. Chitty, L.S. & Lo, Y.M. Noninvasive prenatal screening for genetic diseases using massively parallel sequencing of maternal plasma DNA. *Cold Spring Harb. Perspect. Med.* **5**, a023085 (2015).
22. Tsui, N.B. *et al.* Noninvasive prenatal diagnosis of hemophilia by microfluidics digital PCR analysis of maternal plasma DNA. *Blood* **117**, 3684–3691 (2011).
23. Li, M. *et al.* Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.* **87**, 237–249 (2010).
24. He, Y. *et al.* Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**, 610–614 (2010).
25. Eastman, P.S. *et al.* Maternal viral genotypic zidovudine resistance and infrequent failure of zidovudine therapy to prevent perinatal transmission of human immunodeficiency virus type 1 in pediatric AIDS Clinical Trials Group Protocol 076. *J. Infect. Dis.* **177**, 557–564 (1998).
26. McMahon, M.A. *et al.* The HBV drug entecavir - effects on HIV-1 replication and resistance. *N. Engl. J. Med.* **356**, 2614–2621 (2007).
27. Snyder, T.M., Khush, K.K., Valantine, H.A. & Quake, S.R. Universal noninvasive detection of solid organ transplant rejection. *Proc. Natl. Acad. Sci. USA* **108**, 6229–6234 (2011).
28. Kukita, Y. *et al.* High-fidelity target sequencing of individual molecules identified using barcode sequences: *de novo* detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients. *DNA Res.* **22**, 269–277 (2015).
29. Gregory, M.T. *et al.* Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic Acids Res.* **44**, e22 (2016).
30. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
31. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–655 (2014).
32. Mouliere, F. *et al.* High fragmentation characterizes tumour-derived circulating DNA. *PLoS One* **6**, e23418 (2011).