OligoSeq: Rapid nanopore-sequencing of single-stranded oligonucleotides

Robert Ramirez-Garcia[1*] **and Thomas Heinis**[1*]

[1]Department of Computing, Faculty of Engineering, Imperial College London, London, SW7 2RH GB, United Kingdom

Nanopore-based DNA sequencing technology has achieved remarkable success in sequencing increasingly long DNA strands for genomics research and biotechnology applications. However, the same level of progress has not been achieved for shorter sequences, such as RNA transcripts and DNA oligonucleotides. Oligonucleotides play a crucial role in automated genome engineering through oligo library generation and in DNA data storage, where they encode binary information in DNA libraries. The accurate sequencing of large and diverse oligonucleotide libraries hence is essential, however, accessible sequencing solutions for oligonucleotides — particularly DNA primers for PCR, oligo DNA libraries for mutagenesis or RNA libraries used in gene expression analysis — remain inadequate. Current biological research platforms still rely on expensive and cumbersome quality control methods such as high-performance liquid chromatography and mass spectrometry. To address this gap, we develop OligoSeq, an innovative approach that integrates two complementary techniques: AmpliSeq (based on PCR) and RevSeq (based on reverse complementation of either sequence-specific or random primers) to facilitate sequencing of single-stranded oligonucleotides. OligoSeq uses nanopore technology and can be used as a reference for other sequencing platforms requiring double-stranded adapters, offering a practical and scalable alternative for standard quality control in single-stranded oligonucleotide synthesis.

DNA sequencing | oligonucleotide sequencing | OligoSeq | AmpliSeq | RevSeq | random primers | nanopore technology

*Correspondence:
Robert Ramirez-Garcia: r.ramirez-garcia@imperial.ac.uk
Thomas Heinis: t.heinis@imperial.ac.uk

DRAFT

Introduction

DNA sequencing technology has made significant advances in sequencing long DNA strands for genomics and biotechnological applications. However, similar progress has not been achieved for shorter sequences, such as RNA transcripts or DNA oligonucleotides. Oligonucleotides play a crucial role in synthetic biology, enabling the creation of extensive libraries for automated genome engineering [1][2] and serving as a medium for encoding binary data in DNA storage [3]. This underscores the importance of accurately sequencing large, diverse, and dynamically varying oligonucleotide libraries. However, oligonucleotide sequencing methods, including those for RNA libraries used in gene expression analysis and protein bioproduction optimization [4], remain neither readily accessible nor standardised. Instead, quality control often depends on expensive techniques such as High-Pressure Liquid Chromatography (HPLC) and Mass Spectrometry (MS) [5][6][7]. *OligoSeq*, a method comprising both `AmpliSeq` and `RevSeq`, offers a promising solution for rapid, accessible, and standardised sequencing of DNA and RNA oligonucleotides using nanopore technology. Moreover, it is adaptable to other sequencing platforms that require flanking adaptors for the analysis of double-stranded nucleic acids. Many widely available DNA sequencing technologies require double-stranded DNA (dsDNA) as a foundational substrate for sample preparation and sequencing. This requirement is seen across a range of platforms, including standalone technologies like PacBio [8], and adapter-mediated methods such as Oxford Nanopore's nanopore sequencing platforms (MinION, GridION, and PromethION) [9]. Illumina sequencing, which involves A-tailing and blunt-end adapter binding, also relies on dsDNA for efficient sequencing [10]. Because most sequencing platforms use dsDNA as substrate for sequencing, the demonstrated protocols for the systematic conversion of ssDNA to dsDNA are extendable to these other sequencing platforms.

The use of single-stranded DNA (ssDNA) oligonucleotides, despite not being as popular in biological experimentation where genomics samples, usually in the order of many-thousands of base pairs (bp) in size are commonplace, it is still an essential part of the workflows in a molecular biology laboratory, and it is thus of value for its use in biotechnology, such as in PCR[11], DNA libraries[12], DNA origami[13], and even for DNA data storage[3]. The possibility of sequencing oligonucleotides in a rapid, accessible and standardised, while still being cost-effective and straightforward will be a significant achievement in the facilitation of quality control of oligonucleotides. In this light, different strategies are proposed that use both already established and customised techniques for converting ssDNA oligonucleotides into dsDNA, to repurpose currently available and accessible DNA sequencing technology such as Oxford Nanopore Technologies (ONT). ONT technology is used here as a reference sequencing framework, where dsDNA, compatible with ONT adapters, is required for sequencing.

The proposed strategies, which range from PCR amplification, reverse complementation and *random priming* complementation of ssDNA for double-stranding to dsDNA (also referred to as second-strand synthesis) are demonstrated as a means to make an original sample of short ssDNA oligonucleotides into a dsDNA ready for ONT

sequencing compatibility. These methods are then compared to the direct sequencing of ssDNA, a negative control incompatible with the binding of ONT adapters and thus unable to yield reads for sequence matching (Figure 3).

The number of anchor matching reads with capacity to align (*i.e.*, sequencing output sequences that can be aligned to the target, or query, intended sequence, see Figure 3) increases significantly when using strategies which make the DNA double-stranded using sequence-specific primer sequences. Therefore, the use of known primer sequences for reverse complementation or PCR-amplification of ssDNA, here termed as `AmpliSeq` and `RevSeq`, (which can be readily extended to RNA sample preparations, despite not shown in this work) into its double-stranded homologous, previous to sequencing processing, yields the best sequencing outputs, *i.e.*, with the most number of `anchor matches` (Figure 3) that have capacity to positively contribute to meaningful sequence elucidation to the target oligo, or oligo library, sequence able to produce a consensus sequence that is more or less at a pair with its reference (Figure 4). In this work, `RevSeq` is proposed as superior to `AmpliSeq`, despite a decrease in the number of `anchor matches` is observed, because PCR bias may be a concern for some studies and applications where sequence diversification and variability is at stake.

Overall, the strategies proposed in this work seek to reduce sample processing compared to HPLC/MS while also streamlining the workflow for reverse complementation. By using N9 primers (*N9-Klen* in Figures 3 and 4), our approach eliminates the need for known primers in reverse complementation, which is typically required when using high-temperature polymerases. This aids in the design of simplified pipelines for automation of standardised protocols for studying different samples of ssDNA, which will help, *e.g.*, with the analysis of stock libraries for the quality control of ssDNA in a rapid and cost-effective manner. This work presents various strategies for DNA sequencing, with the potential to be extended to RNA oligonucleotides using adapter-based technologies such as ONT.

Additionally, these protocols could be adapted for PacBio and Illumina, both of which require a double-stranded substrate for sequencing. By leveraging these approaches, the study eliminates the need for costly quality control methods like HPLC (High-Pressure Liquid Chromatography) or MS (Mass Spectrometry). A key advantage of *OligoSeq* is its ability to provide a rapid, accessible, and comprehensive quality assessment of single-stranded oligonucleotide samples using reverse complementation (Figure 4).

This approach enables evaluation of both read quantity and quality, with read frequencies serving as a reliable guide—a claim validated through two independent methods, `RevSeq` and `N9-Klen` (a modality of `RevSeq` because it relies on reverse-complementation or second-strand synthesis), as demonstrated in Figures 3 and 4. The results obtained highlight that the different methods showcased have different advantages and inconveniences, but that the best approaches necessarily entail the pre-processing of the single-stranded oligonucleotide into its double-stranded homologue, by either amplification (`AmpliSeq`), reverse complementation (`RevSeq`) (Materials and methods M5), or random primer second strand synthesis using nonamers (`N9`) (Materials and methods M4), prior to sample preparation for sequencing (see Materials and methods M2). The techniques and strategies presented in this work are intended to serve as a valuable reference for both field biologists developing cost-effective sequencing platforms to sequence RNA (*e.g.*, for transcriptomics analysis) or ssDNA libraries, *e.g.*, prior to standard cloning, mutant library generation or for high-throughput genome engineering.

Conversion of single-stranded DNA into double-stranded DNA. There are several means to convert single-stranded oligonucleotides (such as DNA and RNA) into its double-stranded homologue. The omnipresent Polymerase Chain Reaction (PCR)[14][15] is the most readily accessible technique for double-stranding DNA: PCR is a standard laboratory technique that can be used to "amplify" (`AmpliSeq`) or modified to "reverse complement" (RevSeq) short fragments of DNA, including both dsDNA and ssDNA, or RNA, oligonucleotides.

AmpliSeq: simple PCR-amplification for double-stranding oligonucleotides. The amplification of single-stranded oligonucleotides is possible through the use of short flanking DNA sequences, called oligonucleotides of approximately 20 nucleobases long, that can be used as a "primer" sequences for direct amplification of DNA by means of a DNA polymerase.

Figure 1 demonstrates the basic principle of how two primers (`FW`, *forward*; and `RV`, *reverse*) can be used to direct a DNA polymerase to exponentially amplify a single-strand of DNA after at least a single initial step of reverse complementation (*i.e.*, Figure 1 shows how the $\mathtt{RV}_{20}$ primer is first double-stranding ssDNA into dsDNA followed by the subsequent cycles of a standard PCR undergo). Thus, a PCR amplifies the single-stranded DNA strands that have been primed.

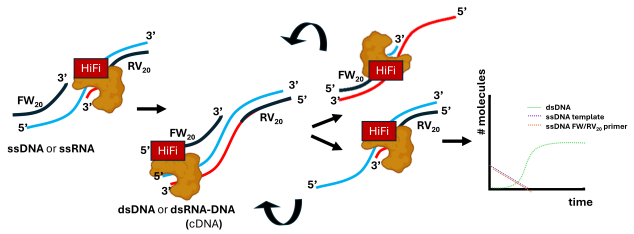

Fig. 1. Using standard forward ($\mathtt{FW}_{20}$) and reverse ($\mathtt{RV}_{20}$) primers for exponential amplification of a single-stranded DNA template using PCR. The `HiFi` stands for High-Fidelity, high-temperature, DNA polymerase.

While the PCR-amplification procedure shown in Figure

1 above is the most readily accessible and effective (as reflected by the number of anchor matches relative to the other analysed samples in Figure 3) means by which to convert ssDNA into dsDNA, it has the disadvantage that PCR bias is not the most suitable approach to representatively sequence the broad sequence space of *e.g.*, an DNA oligo library with diversified sequences or point mutations. Likewise, if the goal of *OligoSeq* is the quality control of an oligo sample supposedly having a singular DNA sequence, PCR (`AmpliSeq`) would not be able to representatively capture synthesis errors, degradation, and other quality issues.

In this light, PCR bias refers to the tendency of a PCR-amplified DNA library to lose its sequence diversity during successive amplification cycles, leading to the overrepresentation of certain sequences at the expense of others [16]. This occurs because, in repeated PCR cycles, DNA strands that are amplified earlier tend to become progressively enriched, resulting in their disproportionate dominance within the library. A key strategy to mitigate bias when regenerating DNA libraries is the use of emulsion PCR (ePCR) [11]. This technique, widely utilised in biological research, enables the amplification of a DNA library while preserving sequence diversity. By encapsulating individual DNA molecules within isolated picoliter-scale water-in-oil droplets, ePCR ensures that each sequence undergoes independent amplification, preventing the overrepresentation of subset sequences and maintaining the homogeneity of all possible DNA amplicons in an oligonucleotide library.

RevSeq: Reverse-complementation techniques for double-stranding oligonucleotides. Reverse-complementation techniques are based on the idea that single-stranded DNA (ssDNA) can be reverse-complemented with a complementary second-strand to form a double-stranded DNA (dsDNA) equivalent that can be suitable for DNA sequencing, such as ONT, but without exponential amplification, and amplification bias, characteristic of PCR; due to only one reverse (`RV`) primer present during the polymerization. The absence of a forward (`FW`) primer in the reaction mixture prevents exponential amplification of ssDNA strands. As a result, only a quantity of dsDNA molecules equivalent to the initial ssDNA substrate remains after reverse complementation, as illustrated in Figure 2. A complementary strand of DNA can be either synthesised using commercially available DNA synthesisers, or by complementing the ssDNA by means of a DNA polymerase (which in this work can be either Q5 (Q5® High-Fidelity 2X Master Mix, NEB, UK) or the Klenow fragment (DNA Polymerase I, Large Klenow Fragment, NEB, UK); and a primer complementary to the reverse flanking end of the complemented DNA sequence (depicted as $\mathtt{RV}_{20}$, Figure 2), or a random nonamer (N9) randomly complementary to the reverse-complemented ssDNA (depicted as $\mathtt{N}_9$).

There are several advantages on the use of DNA polymerases to generate the reverse-complementary strand of the target oligonucleotide over the use of gene synthesis methods. The most evident advantage is cost-effectiveness. A 75-nucleobases long oligonucleotide can cost in the order of £20 using an external provider (IDT, US) or up to £40 if using an *in-house* oligo synthesiser such as Kilobaser (Kilobaser GmbH, Austria). The cost doubles if a reverse-complementary strand is needed. Instead, using a DNA polymerase, such as Q5, Phusion or DNA polymerase I (NEB, UK) the cost can go down to £0.95 per reaction (according to current NEB prices for the reagents involved).

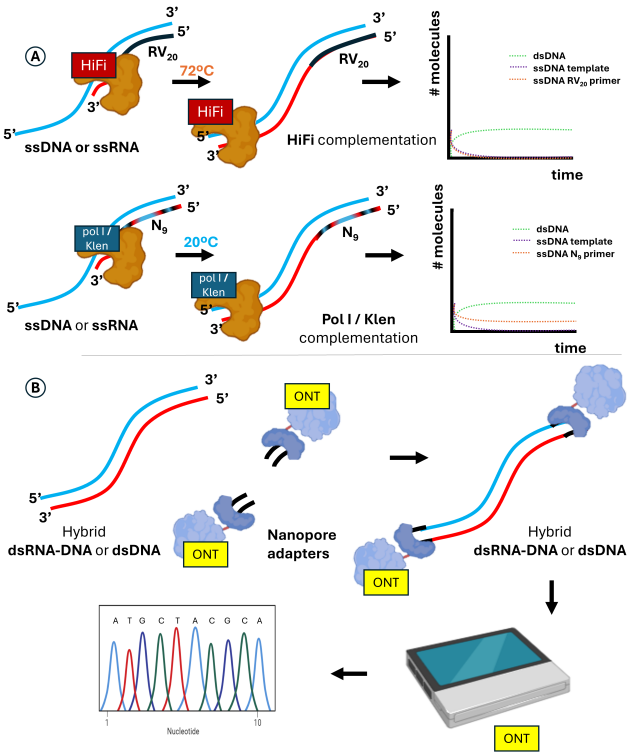

Fig. 2. **A)** Using standard reverse primers ($\mathtt{RV}_{20}$), versus a random primer (nonamers, $\mathtt{N}_9$) approach, for double-stranding complementation of ssDNA molecules enables **B)** ONT-adaptor compatibility and makes the double-stranded substrate suitable for ONT nanopore sequencing readout. `HiFi` stands for High-Fidelity, high-temperature, DNA polymerase (Q5 from NEB, UK) while `PolI/Klen` stands for low-temperature Klenow fragment of the DNA polymerase I.

Second strand synthesis using high-temperature high-fidelity polymerases. The use of high-fidelity (`HiFi`) polymerases, has the advantage that, by using a known primer sequence of about the standard length of 20-nucleobases, it is possible to reverse-complement oligonucleotides (as depicted in Figure 2) with a representative sample of the errors contained among the synthesised DNA strands (*i.e.* avoids PCR bias, as discussed previously).

Second-strand synthesis using random nonamers. The use of short, random, primers of only 9 nucleobases-long (nonamers) presents itself as a opportunity to mitigate the need to have available a particular primer sequence for the sequence-specific reverse-complementation. Thus, the use of random primers (here nonamers) presents itself as a potential alternative to non-specifically reverse-complement "everything within a DNA library".

However, the use of random nonamers imposes temperature constraints during reverse complementation, as the priming oligo's length directly influences its melting and annealing temperatures. Standard PCR reagents, which rely on high-temperature polymerases, are not well suited for this process. As a result, effective enzymatic reverse-complementation requires a polymerase that can function efficiently at lower temperatures. Consequently, standard *HiFi* polymerases commonly used in laboratory PCR are not suitable for applications involving small, random primers. Fortunately, the native DNA polymerase I of *Escherichia coli* is an enzyme that naturally operates at low temperatures of 37 ºC, at its peak performance, and that are suitable for performing at room temperatures (*e.g.*, 20 ºC on a thermocycler). The protocol used for reverse-complementation is described in Materials and methods M4 and M5.

Results

Figure 3 presents a comparative analysis of sequencing read counts of the different samples processed using `AmpliSeq` (PCR), or `RevSeq` with either a reverse complementing primer (using a Q5 polymerase) or with the random nonamer (N9) approach (using the Klenow Fragment of the DNA polymerase I). For each sample, the total number of sequencing reads (white bars) and the number of reads that passed the anchor filter (black bars), give a measure of the proportion of reads within the total sequencing sample that can be used to sequence an oligo sample, as a way to assess the quality of the procedure to generate a sequencing-able dsDNA (out of the original ssDNA) material of enough quality and representativeness.

As shown in Figure 3, there is substantial variation, across all samples, in both the total number of reads and the proportion of reads that successfully match the anchor criteria (of 20 nt at the 5' ends of the plus `(+)` and minus `(-)` DNA sequences, respective to a plus `(+)` reference sequence, see Script S1). In several samples, such as the `75 nt AmpliSeq` and `150 nt AmpliSeq Q5 (Kilobaser)` and the `150 nt RevSeq Q5 (Kilobaser - Replica)`, a high proportion of reads (ranging from approximately 52% to 70%) pass the anchor filter (see Materials and methods M3), indicating efficient and specific amplification (or reverse complementation in case of `150 nt RevSeq Q5 (Kilobaser - Replica)`) or sequencing of the target region. This suggests that the use of both PCR and reverse complementation with defined primers using a Q5 polymerase was effective in generating double-stranded reads that align well with the reference anchors (without regard to the rest of the sequence, which is analysed in Figure 4).

Conversely, other samples, such as the `75 nt` and `90 nt N9 Klenow (Kilobaser)`, show a much lower proportion of anchor-matched reads (as low as 0.6% to 1.6%), despite having a moderate total read count (*i.e.* all samples were normalised to 200 fmol as evidenced with the consistence of the total reads accross samples in Figure 3). This low efficiency could reflect issues such as non-specific reverse complementation or amplification using random nonamers (`N9`), or a high background of off-target sequences due to the origin of the single-stranded DNA. However, because when using quality samples synthesised with IDT (see `75 nt` and `90 nt 150 nt N9 Klenow (IDT)` gave an over two-fold (logarithmic) improvement in the number of anchor matches (Figure 3) with values of 8.7%, 17.8% and 14%, respectively, it is fair to suggest that quality of the original sample of DNA, rather than unspecific or off-target amplification, is an observable feature of the results obtained with an *in-house* synthesised single-stranded DNA (Kilobaser) versus an industry-standard single-stranded DNA (IDT).

As expected, with the `90 nt ssDNA Control (IDT)` (used as a negative control), no reads passed the anchor filter because no double-stranding of single-stranded DNA was realised.

Intermediate results were observed in samples like the `75 nt` and `90 nt RevSeq Q5 (IDT)`, where approximately 39–49% of reads are anchor-matched, which may suggest partial success in targeting the desired region, but the authors attribute it to the capacity of `RevSeq` to capture the variability of the oligonucleotide sample to a much greater degree than the `AmpliSeq` samples, because `RevSeq` avoids PCR bias and therefore is able to show greater variability within a sample sequence (as depicted in Figure 4).

The `dsDNA annealed noPol (IDT)` samples (for `75 nt` and `90 nt`, respectively) were used here as a positive control in which no enzymatic processing of single-stranded DNA was realised and instead the reverse complementing DNA strand was provided and annealed to the leading strand as a means to obtain the dsDNA intermediate for ONT sequencing sample preparation (Materials and methods M2), using industry-grade DNA from IDT. This positive control shows that, when using the dsDNA-annealed IDT samples, moderate anchor-matching rates of up to 41–48% can be expected, indicating that a significant fraction of the reads correspond to the intended sequence. These controls also establish that when using ONT, it is expectable to see a consistency of over 50% of total sequenced reads not being meaningful to the reference sequence. Surprisingly, synthesising both strands at industry quality grade did not yield better performance (measured as high `anchor matches`, Figure 3, and low frequency variability, Figure 4) than the observed with`AmpliSeq`, further reinforcing the idea the synthesis quality can be evidenced using this protocol, and that the remarkably better results of the `AmpliSeq` homologue samples, are likely due to PCR bias and the enrichment of meaningful reads due to amplification of a more homogeneous anchor matching

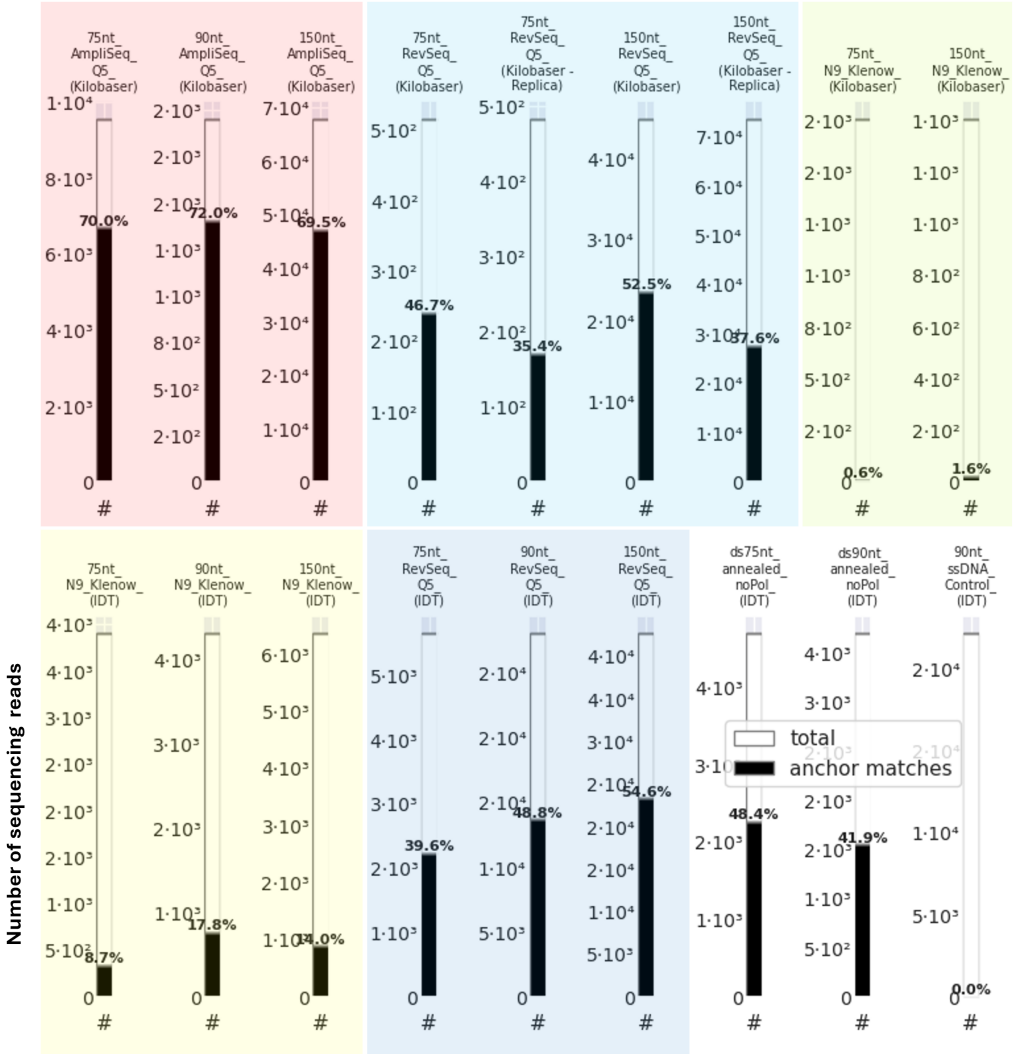

Fig. 3. Number (#) of sequence anchor matches that can be aligned to target sequence. Total: Total number of reads extracted from the sequencing run. **Anchor matches:** Sequences that could be aligned to a reference sequence through anchoring reads to the first 5' end 20 nt of the plus `(+)` and minus `(-)` DNA sequences filtered using Script S1, regardless of % identity beyond the 20 nt anchors. All experiments were normalised to 200 fmol per load of dsDNA (`AmpliSeq` and `RevSeq`, either with Q5 or N9-Klen) or ssDNA (Control). DNA was synthesised using an *in-house* Kilobaser DNA Express (Kilobaser Gmbh, Austria) or Integrated DNA Technologies (IDT, US).

sequence pool, which does not occur with the `RevSeq` samples, nor with the directly annealed dsDNA samples, because these better reflect the variability of sequences implicit of samples with high, but still limited to a certain extent, synthesis quality.

Overall, these results demonstrate the effectiveness of the methods proposed for the rapid sequencing and quality control of single-stranded oligonucleotides using nanopore sequencing and highlight the importance of sequencing depth and the specificity of read alignment to the reference anchors (*e.g.* high total read counts are only valuable if a substantial proportion of those reads are relevant to the target region). The anchor-matching percentage (Figure 3) serves thus as a key quality metric, reflecting the specificity (and bias) of the experimental protocol for each sample (*i.e.* accounting for `AmpliSeq` or `RevSeq`, specificity of primers and polymerases). Samples with high anchor-matching rates are likely to indicate a high-quality of original ssDNA oligo material (as corroborated with Figure 4) and promise a more reliable substrate for downstream applications, while those with low rates may indicate a low quality material or reflect a high variability in sequence, which may still be a positive feature if *e.g.*, building libraries for specific downstream applications such as mutagenesis and other diversification purposes.

Figure 3 shows the number of reads for "`total`" (0–100% matching) and "target 20 nt-forward and -reverse `anchor matches`" used to determine the quality of the oligonucleotides that have been prepared for ONT sequencing using the sequencing techniques showcased in this work. The "`total`" number of sequences (Figure 3) belongs to sequencing reads without regard to whether or not these can be aligned to the reference. Conversely, the "`anchor matches`" reads belong to the sequencing reads that can be

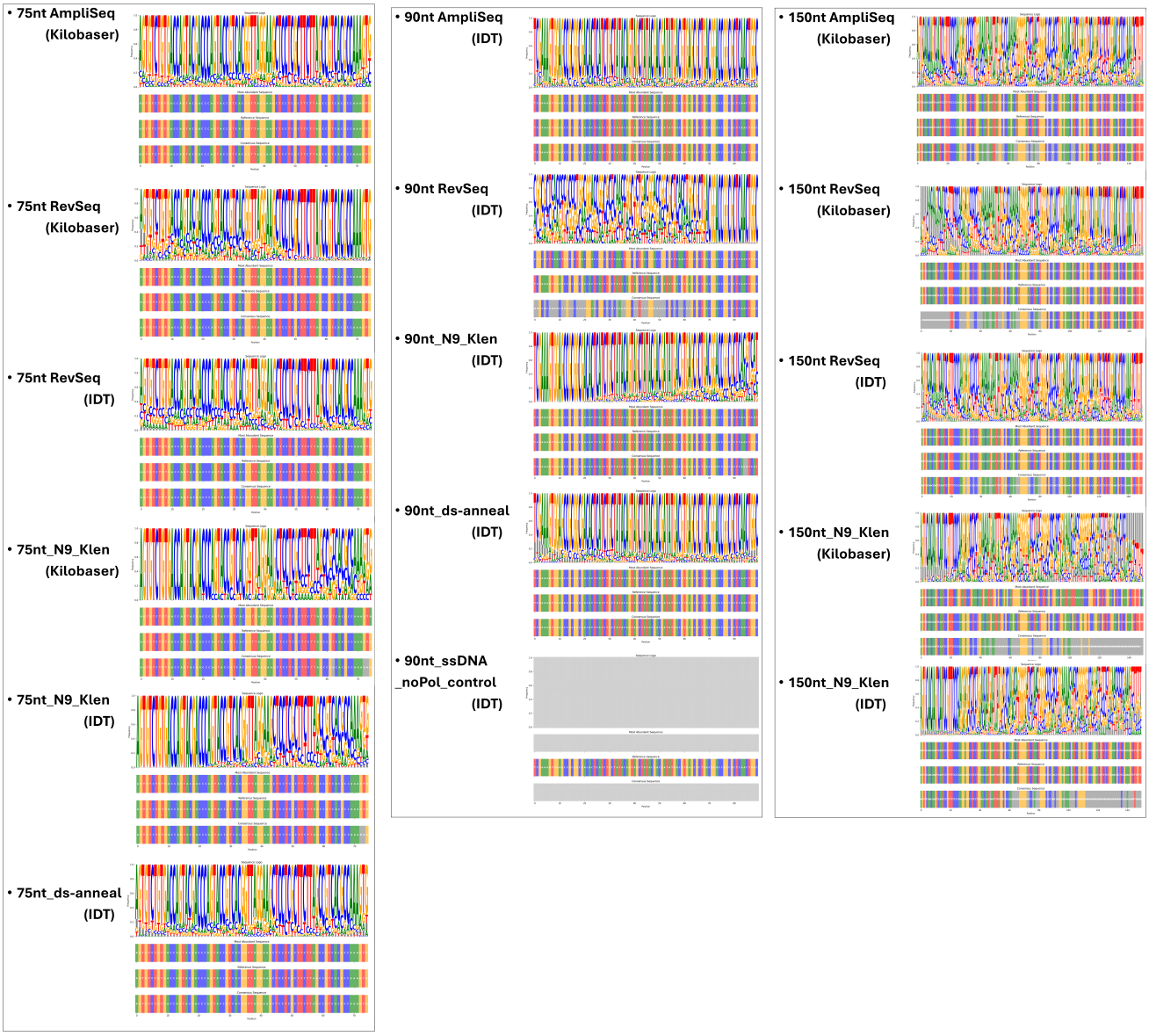

Fig. 4. Sequence logo plots for the assessment of oligo quality per sample preparation method. The abundance of a nucleobase at a particular position is proportional to the size of the nucleobase logo in the plot, and the variability is thus represented as a conjunction of sequence logos that can be used to determine a **consensus sequence:** The overall sequence considering all variant reads; and the **most abundant sequence:** which corresponds to the most abundant sequence (and therefore the most likely) after alignment of all reads, when matched against a **reference sequence**.

aligned (because of the anchor match), with no penalising algorithms, to the reference sequence used for synthesis and therefore are "read" candidates that will contribute to the sequence logo (Figure 4), representative of the real content of the sequenced oligonucleotide sample.

Figure 4 presents sequence logo plots for each sample, illustrating the nucleotide frequency at every position along the synthesised DNA oligonucleotides. These plots provide a detailed view of sequence conservation and variability across the different synthesis and reverse complementation strategies. Each logo comes with a colour-coded representation of the most abundant sequence, the reference sequence, and the consensus sequence, to facilitate a direct visual comparison. Across most samples, the sequence logos reveal high conservation at the anchor regions, as expected from the anchor-filtering approach described in Figure 3. However, it is important to gighlight that the nature of alignment, because it starts with the sequences that were anchored also conditions a higher conservation of sequences at the anchor sites with the most abundant sequences. This is reflected by tall, sharply defined letters at the sequence ends (Figure 4), indicating that the majority of reads align precisely to the reference at these positions. In contrast, the more central regions of the sequences may display varying degrees of nucleotide diversity, with some samples showing a more pronounced sequence variability as evidence of mutations or synthesis errors. For example, samples synthesised using

the N9 Klenow or ds-anneal approaches (notably those from IDT) show increased sequence heterogeneity, as indicated by shorter, more mixed stacks of letters in the logo plots. This suggests a higher frequency of incorporation errors or incomplete synthesis, particularly in the central regions of the analysed oligonucleotides.

In regards to the frequency of deletions displayed in Figure 4, it is important to justify the fact that deletions at the ends of the sequences do not necessarily imply an accumulated frequency of deletions at the end positions, but that these deletions can accumulate anywhere in the sequence and then evidenced as deletions at the end-sequence due to the shortening of the aligned sequences to the reference sequence (which in Figure 4 are always aligned without penalisation for deletions, substitutions or insertions due to the purpose of the plots simply being the evidence of sequence variability, rather than alignment of sequences to a reference). Therefore, the observed deletions at the end-sequences contribute to the frequencies observed at each position throughout the entire sequence, as a result of shortened sequences in case of point deletions (at any point of the sequence) that produce the frame shifts, e.g., `150nt N9 Klen`.

Therefore, these "shortening" of sequences due to synthesis problems (*i.e.* insertions, deletions or substitutions) can be attributed to both, deletions at any position within the sequence, and termination of synthesis due to extended lengths (which is more patent when using Kilobaser 150-mers, respective to their IDT homologues). This observed sequence variability due to termination of synthesis is also consistent with the chemical synthesis process, which proceeds from the 3' to 5' end of the plus `(+)` strand. Errors and mutations tend to accumulate toward the 5' end, as each successive nucleotide addition carries a risk of incomplete coupling or side reactions. Additionally, the subsequent reverse complementation, performed using various enzymatic strategies, further influences the fidelity of the final double-stranded product. In this regard, samples prepared with the high-fidelity polymerase (Q5) and biasing protocols (*i.e.*, `AmpliSeq`) generally display lower sequence variability and fewer mutations, as evidenced by the dominance of the reference base at each position in the logo plots. In contrast, samples subjected to less stringent, more diversely representative, or more error-prone reverse complementation `e.g.` using either the `RevSeq` with normal reverse-complementing primers, or a random nonamer (N9) for the Klenow fragment of the DNA polymerase I, show greater sequence diversity, with some positions showing no clear consensus base. In the same line, sequences synthesised using industry grade DNA, particularly at DNA lengths longer than 75 nt also display improved metrics (i.e. a decreased diversity of sequences and a more consensus sequence).

Overall, and despite minor differences that can be attributed to PCR bias, substrate DNA length and quality of the original DNA sample, it is possible to observe that all studied cases are able to retrieve fully matching sequences (with degree of variability implicit to sample quality), and where consensus is lesser (particularly observed at DNA sequences of 150 nt) it is still possible to retrieve the original sequence using the "Most Abundant Sequence" based on the most frequent aligned sequences retrieved from the sequence logo of Figure 4. These findings are in agreement with the anchor-matching results shown in Figure 3, where samples with high anchor-matching rates also tend to show greater sequence fidelity throughout the oligonucleotide. Conversely, samples with low anchor-matching rates often display increased sequence heterogeneity and a higher incidence of mutations, reflecting both synthesis and processing inefficiencies. Notably, the `90 nt ssDNA noPol control` sample shows no meaningful sequence logo, consistent with the absence of anchor-matched reads in this condition.

Overall, the sequence logo analysis underscores the impact of both chemical synthesis and enzymatic processing on the quality and fidelity of synthetic DNA oligonucleotides. The combination of anchor-filtered read selection and sequence logo visualization provides a useful approach for assessing sequence integrity, identifying sources of variability, and optimizing protocols for high-fidelity DNA synthesis and processing.

The analysis of Figure 3 shows that `AmpliSeq` is the method that yields the most homogeneous sample, with the highest proportion of aligning reads that fulfil the proposed cut-off anchor sequences of 20 nt at both 5' and 3' ends of the reference sequence (Figure 4). This is consistent with the observation that PCR can enrich the sample of ssDNA into an exponentially amplified dsDNA sample, but also of PCR bias. Instead, the results obtained for the `RevSeq` and N_9 experiments *i.e.*, either with `HiFi` Q5 polymerase, or the Klenow fragment of DNA polymerse I, show much lower number of target anchor reads, which belong to the fact that reverse-complementation does not amplify the sample, but simply produces the complementary strand of the leading strand of ssDNA, thus producing samples with similar number of molecules at the input than at the pre-processing output (*i.e.*, input ssDNA vs dsDNA output), which according to the positive controls, should be no higher tan 50% (Figure 3). However, within reverse-complemented samples, a consistent, orders of magnitude, decrease in the relation of the number of total and target anchor match reads, between all samples is obtained. An increase in the variability of the anchor matching target sequences from the oligo sample is also observed, respective to results obtained with the PCR-based `AmpliSeq` experiments, which is consistent with the hypothesis of PCR bias mentioned in the introduction, and the fact that the nature of reverse-complementation does circumvent it.

The two different methods for reverse-complementation used, RV_{20} and N_9, show that similar variability profiles (as

demonstrated in Figure 4) are obtained, despite a remarkably lower proportion between anchor matches and total reads, with the advantage that using N_9 random nonamers facilitate not having to continuously pre-design and re-design primer sequence and give room to a more simplified protocol that is more amenable to automation and high-throughput processing.

In this sense, RevSeq is presented here as a useful method that enables the sequencing of input samples of ssDNA that can be heterogeneous and therefore have a greater variability, needed in experimental set ups where large and diverse DNA and RNA libraries are used (*e.g.* a library of mutants), through a protocol that is accessible, rapid, and standardisable, using either HiFi polymerases with standard, 20-nucleobase long, primers, or the use of N_9 primers (Figure 2) as an interesting alternative for quality control of oligonucleotide samples.

The RevSeq approach, with its RV_{20} or N_9 variants, demonstrated has the advantage that the PCR amplification bias can be circumvented, despite at the expense of a lower number of anchor match reads (Figure 3) that can be aligned to a reference sequence based on 20 nt anchors (at both sides of the reference sequence, Figure 4), and is therefore presented as a suitable approach for determining variability within a sample library of oligonucleotide sequences which, as aforementioned, could comprise diversified point mutations, due to randomization, variant selection, or oligo synthesis quality issues.

Using the results demonstrated in Figure 4 it is easy to see that when using a challenging sample e.g., a Kilobaser synthesis of a 150-mer (150nt_AmpliSeq, 150nt_RevSeq and 150nt_N9_Klen), the oligonucleotides analysed display a greater diversity and a lack of consensus, which although does not prevent to retrieve the original sequence using "the most abundant reads", exemplify the limitations of synthesis of long oligonucleotides (including industry-grade standard 150-mers from IDT) compared to more feasible sequence lengths (75-mers and 90-mers) also tested in this work. In this work, a cut-off sequence of 75 nt-long was used, taking into account that, for long, industry standard single-stranded oligonucleotides have been capped to 80–90 nucleobases.

In this line, for long 150-mers, the origin of the oligo sample decreases significantly in the Kilobaser synthesised oligos, respective to the IDT synthesised ones, despite the variability (upper frequency plots in Figure 4) shows that always the most likely sequence matches that of the query reference used for synthesis in all cases. This should of course be attributed to the differences in sample quality that come from a different procedure for synthesis and purification and the quality control checks in place, which are patent in IDT samples, but not in our procedure for oligo synthesis using Kilobaser.

Remarkably, for 150-mers where full DNA lengths are more difficult to achieve (synthesis-wise), the general increase in frequencies usually show a slight increase in the variability of sequences a the 5-end of the sequenced oligonucleotides, which translates into a consensus sequence position labelled as "N", which is consistent with the fact that sequences are synthesised 3' end to 5' end. However, errors at the sides of each frequency plot should also be attributable to the nature of the alignment that always goes from the anchor to the rest of the sequence, prioritising better reads early in the alignment. To mitigate this artifact, both references (plus (+) and minus (−) strands) are used for the alignment (see Script S1 and Figure 5).

In summary, from the analysis of the results demonstrated in Figure 4, when using an oligonucleotide sample synthesised by IDT (Integrated DNA Technologies, US) *i.e.*, the 90nt_RevSeq (using Q5 and a RV_{20} primer) vs 90nt_N9_Klen (using the Klenow fragment of the DNA polymerase I and a N_9 primer) show consistent results (Figure 4) in regard to the variability displayed for the different reads of the sample. This highlights that both RevSeq strategies (RV_{20} vs. N_9) are equally valid and demonstrate a similar variability which cannot but be correlated to the real diversity of sequences within the analysed sample. The results also demonstrate that what is described for the 90 nt sequence (synthesised and quality-guaranteed by IDT) does also apply to the *in-house* Kilobaser-synthesised samples of 75 nt and 150 nt, respectively, which display similar profiles for the two methods (RV_{20} vs. N_9) of RevSeq used. This is also indicative that the improvement in the number of reads; but more importantly, the increased (although not remarkably) sequencing quality (as exemplified by a less diverse sample and a more defined consensus sequence, Figure 4) of the samples processed using AmpliSeq (*i.e.* PCR) is likely to be due to PCR bias.

In regards to the 90 nt sequence (IDT), it is observed, that for the cases of reverse complementation, the yield of high identity matching sequences was slightly lower (Figure 3, without a compromise in the final sequences retrieved through "the most abundant sequence" shown with its variability (Figure 4). This can be attributed to the fact that the 90 nt sequence showcased, comes from an IDT tube that over-a-year-old and that has been frozen and thawed multiple times throughout this year, while all other samples where immediately processed after synthesis. This highlights the more how the techniques shown in this work are of value to assess the quality of single-stranded oligonucleotide samples, which can be compromised by synthesis technique but also by storage conditions and time.

Ultimately, Figure 4 demonstrates how RevSeq, and specially with the random nonamers (N9) technique, can be of use for preliminary quality control of single-stranded oligonucleotides that span from 75 nt to 150 nt. Of relevance

is the consistency to which the reverse complementation, either using Q5 polymerase or the Klenow fragment with random nonamers, is able to represent the variability of the sample. In this sense, the consistency shown when comparing of Q5 and N9 is of interest because using random primers (*i.e.* N9 nonamers) allows to automate the process of quality control without regard to the sequences to analyse, and with the possibility to do this at room temperatures (*e.g.* 20 ºC) which is useful for field applications outside the laboratory.

The results demonstrate a first-in-kind sequencing and quality control protocol of single-stranded oligonucleotides using a DNA sequencing technology (in this case using ONT, but the sample processing could be certainly adapted for PacBio and Illumina, despite at a larger cost and processing times), as an alternative to more comprehensive (and expensive) methods such as HPLC/MS.

Conclusion

As expected, the results obtained highlight `AmpliSeq` (PCR), and `RevSeq` (reverse-complementation) as methods for the sequencing of single-stranded DNA oligonucleotides of 75 nt, 90 nt, and 150 nt in length. For all cases, a correspondence between an increase in length and a decrease in quality is observed, regardless of whether the oligos were synthesised using Kilobaser or IDT, although the quality was visibly better when using IDT. A decrease in the quality of IDT samples when using oligos that have been stored and re-stored for over a year (with repeated freeze-thaw cycles) is also observed, respective to samples recently synthesised. A correlation between increased length and lower qualities (as corresponded by increased sequence diversity and decreased frequencies) is also observed as a measure that can help define quality which can be found when using `RevSeq` (either with Q5 or N9-Klen), as opposed to the more prone-to-bias `AmpliSeq`.

Overall, the results in Figure 3 and Figure 4 show that both `AmpliSeq` and `RevSeq` are appropriate techniques for the pre-processing of single-stranded oligonucleotides for oligonucleotide sequencing (*OligoSeq*) using ONT that are presented as a rapid and accessible alternative to currently standardised oligonucleotide quality control via HPLC and/or MS; while `RevSeq` is presented as a method that is able to show a greater variability, which increases as the synthesis quality of the oligo sample decreases, demonstrating the potential of `RevSeq` to study sequence diversity within an oligo library for many applications in biotechnology, such as oligo quality control, assessment of the representativeness of a diverse ssDNA pool, or for mRNA sequencing using cDNA libraries, among many others.

Materials and methods

The materials and protocols used for the preparation of this manuscript are stated as follows.

M1. Selection of experimental samples for analysis. Sequences of 75 nt, 90 nt and 150 nt in length were chosen to highlight higher error rates positively correlated to the length of the oligos used in this work. The samples used were synthesised by either a DNA Express machine (Kilobaser Gmbh, Austria) or Integrated DNA Technologies (IDT) as a means to have a control samples with an industry standard quality control (*i.e.* IDT). `AmpliSeq` (PCR), and `RevSeq` (reverse-complementation) of single-stranded DNA oligonucleotides of 75 nt, 90 nt, and 150 nt in length were realised as a means to compare the performance of the different proposed methods with the length of the oligonucleotides used. When using Kilobaser, a correspondence between an increase in length and a decrease in quality should be expected, and a control using oligonucleotides synthesised by IDT is thus provided, showing also a similar correlation between increased length and lower qualities (as corresponded by increased sequence variability and decreased frequencies) and to assess whether an increase in frequency diversity is found when using `RevSeq`, as opposed to the more prone-to-bias `AmpliSeq`. A control run consisting of an unprocessed, industry-grade, quality controlled, single-stranded 90-mer sequence from an external manufactures (IDT, US) was used.

M2. ONT barcoded sample preparation. All DNA oligonucleotide preparations were resuspended in nuclease-free water (for Kilobaser-synthesised samples) or in IDT-buffer (for IDT control samples) and double-stranded using either `AmpliSeq` or `RevSeq` (with either the RV_{20} or the N_9 primers), the resulting amplified or reverse-complemented products were cleaned-up with two consecutive steps of 80% EtOH, after mixing the sample with a 180 %(v/v) ratio of AMPure XP beads (Beckman, US). Next, a DNA dA-tailing procedure was effectuated using NEBNext® Ultra™ II End Repair/dA-Tailing Module (NEB, UK) by mixing 11 μL with 200 fmol of oligonucleotide sample with 1.75 μL of NEBNext® Ultra™ II End Repair/dA-Tailing Buffer and 0.75 μL of NEBNext® Ultra™ II End Repair/dA-Tailing Enzyme Mix, and incubated at 20°C for 5 minutes and 65 °C for an additional 5 minutes before purification again with AMPure XP beads. A 7.5 μL of each end-prepped doubly-stranded sample was mixed with 2.5 μL of a different Native Barcode (NB01-24), 10 μL of Blunt/TA Ligase Master Mix (NEB, UK) before incubation for 20 minutes at room temperature. The each barcoding reactions was stopped with 4 μL of a solution of EDTA (SQK-NBDD114.24 - ONT, UK) before pooling all reactions into a same Eppendorf tube. The resulting pool was mixed with a 40 %(v/v) ratio of AMPure XP beads (Beckman, US) and cleaned-up with two consecutive steps of 80% EtOH before eluting a mixed pool in 35 μL nuclease-free water. A ligation reaction by combining 30 μL of the purified end-repaired and dA-tailed pool of oligonucleotides was mixed with with 10 μL of NEBNext Quick Ligation Reaction Buffer (5X), 5 μL of Native Adapter (SQK-NBDD114.24 - ONT, UK) and 5 μL of Quick T4 Ligase (NEB, UK) incubated for 20 minutes at room temperature.

Fig. 5. Strand-aware alignment and sequence logo generation using anchor-based read filtering. A) Schematic representation of anchor-based alignment: (i) Forward reads are aligned based on matching a predefined left anchor (black), extending into the reference sequence (green) and right anchor (gray); mismatches (mutations) are marked with red crosses (x). (ii) Reverse reads are reverse-complemented and aligned based on matching a right anchor (black, right), proceeding leftwards along the reference. (iii) All reads passing anchor filters—forward and reverse—are aligned and combined into a uniform forward sequence dataset. **B)** Output visualization of the aligned reads: the sequence logo (top) represents nucleotide frequencies at each position; below, the most abundant sequence, reference sequence, and consensus sequence are shown using colour-coded nucleotide highlights for intuitive and direct visual comparison.

After ligation, a successive modified step for AMPure XP cleanup, in which instead of the two steps of 200 μL of 80 % EtOH for paramagnetic-bead clean-up, two-steps of 125 μL of Short Fragment Buffer (SFB) (SQK-NBDD114.24 - ONT, UK) is used, to remove excess adapters. The DNA is then quantified using a suitable method, such as a fluorometer. Finally, the prepared DNA are loaded onto the nanopore flow cell or flongle, according to the manufacturer's instructions, ensuring proper settings on the MinION device.

M3. Strand-aware alignment using anchor-based read filtering for the assessment of oligo quality. The code for strand-aware sequence alignment and logo generation workflow used for the generation of sequence logo frequency plots used in this work for sequence analysis is available in the Supplementary Material SS1 (Script S1).

For strand-aware alignment using anchor-based read filtering after DNA sequencing using an MinION M1kc device (ONT, UK), the sequencing reads, separated by each barcode (SQK-NBDD114.24 - ONT, UK), as `.fastq` files with a Q score equal or greater than 9, were retrieved from the MinION Mk1C (ONT, UK) device. The reads contained within `.fastq` files were then screened against a reference sequence corresponding to either the 75 nt sequence, the 90 nt sequence, or the 150 nt sequence (see `ref_sequence.fa` in the bash command below) using the executed bash command with Script S1:

```
python strand_aware_logo.py
-i fastq_pass/
--reference ref_sequence
--anchor_len 20
--output logo.png
```

Sequencing reads were then processed to generate strand-aware sequence logos relative to the reference sequence to quality check against (see referene sequence in Figure 4). Script S1 gets as input the directory containing `.fastq` files (i.e the totality of sequencing reads), a reference sequence in `.fasta` format, and a user-defined anchor length (20 nt in length in this work) specifying the number of nucleotides at each end of the reference to serve as alignment anchors. The reference sequence was first read from the `.fasta` file, and all sequencing reads were extracted from the provided `.fastq` file-sequencing-containing folder.

To account for strand orientation, each read was compared to the reference sequence in both the forward and reverse complement orientations. Reads were first checked for a match to the left (5' end) anchor of the reference; if a match was found, the read was aligned accordingly. Reads that did not match the left anchor were reverse complemented and then checked for a match to the right (3' end) anchor of the reference. Only reads matching either the left or right anchor were retained for sequence logo frequency analysis. These selected reads were then padded or trimmed as necessary to match the length of the reference sequence, ensuring consistent alignment across all positions.

For the aligned reads, a position-by-position count of each nucleotide (`A`, `C`, `G`, `T`, and `-` gaps) was computed without penalisation for any insertions, deletions or substitutions, with the goal to reflect the reality of the oligo sample. These counts were normalised to generate frequency matrices, which were visualised as sequence logos using the *logomaker* Python library[17]. The resulting logo represents nucleotide variability at each position relative to the reference and is proposed as a measure of quality, taking into account each of the methods used (`AmpliSeq`/PCR or `RevSeq`, with or without random primers). The script also identifies and displays the most abundant sequence, the reference sequence, and a consensus sequence (defined as the most frequent nucleotide at each position, or '`N`' if no nucleotide is dominant). The final output was saved as a `.png` image containing these visualizations, and summary of sequences passing the anchoring filters used for the frequency sequence logo plots (Figure 4 are plotted in Figure

3 which include the number of reads processed, the most abundant sequence, and the consensus sequence.

In short, this pipeline (Script S1, Figure 5) processes sequencing reads by matching forward reads to a user-defined left anchor and reverse-complemented reads to a right anchor, both derived from a reference sequence, which can correspond to the primers used for PCR amplification (`AmpliSeq`) or reverse complementation using $\mathtt{RV}_{20}$ (`RevSeq`). Afterwards, reads that align (with no penalisation algorithms) successfully are trimmed and aligned based on anchor positions to ensure positional consistency. The processed reads are then used to compute a sequence logo that reflects nucleotide frequency at each position.

M4. Second-strand synthesis using the Klenow fragment of the DNA polymerase I. For reverse-complementation (`RevSeq`) using the low-temperature Klenow fragment of the DNA polymerase I, using random nonamers ($\mathtt{N}_9$) (Figure 2) as priming sequences (`N9-Klen` samples, Figures 3 and 4), a sample of 200 fmol of ssDNA was mixed in 5 μL of 10X rCutsmart Buffer (NEB, UK) 1 μL of a 100 μM solution of random nonamers ($\mathtt{N}_9$), 2 μL of a 10 mM solution of dNTPs (NEB, UK) and 38.85 μL of nuclease-free water. The mixture was annealed heating the mixture at 94 ºC and down to 25 ºC over a period of 10 minutes. Afterwards, 0.15 μL of a 50 mM solution of NAD+ (NEB, UK) and 2 μL of the Klenow fragment (exo-) of the DNA polymerase I (M0212S - NEB, UK) at 5,000 U/μL was added to the mixture before incubation at 20 ºC for over an hour.

M5. Amplification or second-strand synthesis using `HiFi` polymerases. For reverse-complementation (`RevSeq`) using $\mathtt{RV}_{20}$ primers (Figure 2) and the Q5 DNA polymerase (NEB, UK), a sample of 200 fmol of ssDNA was mixed in 22.5 μL of nuclease-free water, 2.5 μL of a 10 μM reverse-complementary primer (*i.e.* referred to as the $\mathtt{RV}_{20}$) to the 5-end of the ssDNA sequence, and 25 μL of 2X Q5 Master Mix (NEB, UK). The mixture was incubated following a standard PCR themocycling protocol (as recommended by NEB manufacturing instructions) over 24 cycles. For `AmpliSeq`, a 20.5 μL of nuclease-free water was used to dissolve the 200 fmol ssDNA, and both 2.5 μL of 10 μM of forward (`FW`) and reverse (`RV`) was used instead.

ACKNOWLEDGEMENTS
The authors are grateful to Patrycja Sokolowska from New England Biolabs (NEB) for her thorough support, correspondence and troubleshooting to define the protocols showcased in this manuscript, using NEB reagents.

DECLARATION OF CONFLICT OF INTEREST
The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS
R.R.G. and T.H. conceived the project. R.R.G designed and conducted second-strand synthesis and sequencing experiments, wrote the code for strand-aware alignment and anchor filtering for sequence analysis, analysed the results and wrote the manuscript. All authors revised the manuscript.

DATA AVAILABILITY
All sequencing raw `.fastq` files are available in the GitHub repository: https://github.com/ImperialCollegeLondon/OligoSeq

Bibliography

1. M. C. Needels, D. G. Jones, E. H. Tate, G. L. Heinkel, L. M. Kochersperger, W. J. Dower, R. W. Barrett, and M. A. Gallop. Generation and screening of an oligonucleotide-encoded synthetic peptide library. *Proceedings of the National Academy of Sciences of the United States of America*, 90(22):10700–10704, 1993. doi: 10.1073/pnas.90.22.10700.
2. Mads T. Bonde, Sriram Kosuri, Hans J. Genee, Kira Sarup-Lytzen, George M. Church, Morten O. A. Sommer, and Harris H. Wang. Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.
3. Roman Sokolovskii, Robert Ramirez-Garcia, and Thomas Heinis. Adaptive sampling in nanopore sequencing for PCR-free random access in dna data storage. *bioRxiv*, 2024. doi: https://doi.org/10.1101/2024.11.05.622081.
4. John J. Harrington, Bruce Sherf, Stephen Rundlett, P. David Jackson, Rob Perry, Scott Cain, Christina Leventhal, Mark Thornton, Rakesh Ramachandran, Jessica Whittington, Laura Lerner, Dana Costanzo, Karen McElligott, Sherry Boozer, Robert Mays, Emery Smith, Neil Veloso, Alison Klika, Jennifer Hess, Kevin Cothren, Kalok Lo, Jason Offenbacher, Joel Danzig, and Matt Ducar. Creation of genome-wide protein expression libraries using random activation of gene expression. *Nature Biotechnology*, 19:440–445, 2001. doi: 10.1038/86873.
5. Ute Bahr, Hüseyin Aygün, and Michael Karas. Sequencing of single and double stranded rna oligonucleotides by acid hydrolysis and maldi mass spectrometry. *Analytical Chemistry*, 81(8):2911–2917, March 2009. doi: 10.1021/ac9001872.
6. Soheil Pourshahian and Sergei M. Gryaznov. Sequencing of phosphoramidate oligonucleotides by acid hydrolysis and mass spectrometry. *Analytical Chemistry*, 91(17):10712–10719, August 2019. doi: 10.1021/acs.analchem.9b02278.
7. Mingming Wang, Brian O'Day, Brian Michaels, Jurjus Jurayj, Bao Zhong Cai, and Tao Wei. Sequencing of phosphorodiamidate morpholino oligomers by hydrophilic interaction chromatography coupled to tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 35(10), August 2024. doi: 10.1021/jasms.4b00145. Open Access.
8. Pacific Biosciences. Pacific Biosciences template preparation and sequencing. *© Copyright 2010 - 2014, Pacific Biosciences of California, Inc. All rights reserved*, 2014.
9. Oxford Nanopore Technologies. Ligation sequencing kit V14 (SQK-LSK114) protocol. *Oxford Nanopore Technologies, Oxford*, 2022.
10. Iraad F. Bronner and Michael A. Quail. Best practices for Illumina library preparation. *Current Protocols in Human Genetics*, 102(e86), 2019. doi: 10.1002/cphg.86.
11. Vaishali Verma, Amita Gupta, and Vijay K. Chaudhary. Emulsion PCR made easy. *Biotechniques*, 69(1):421–426, Jul 2020. doi: 10.2144/btn-2019-0161. Epub 2020 Apr 27.
12. A. Currin, N. Swainston, P. J. Day, and D. B. Kell. Speedygenes: Exploiting an improved gene synthesis method for the efficient production of synthetic protein libraries for directed evolution. In R. Hughes, editor, *Synthetic DNA*, volume 1472 of *Methods in Molecular Biology*, pages 69–90. Humana Press, 2017. doi: 10.1007/978-1-4939-6343-0_5.
13. Paul W. K. Rothemund. Folding dna to create nanoscale shapes and patterns. *Nature*, 440: 297–302, Mar 2006. doi: 10.1038/nature04586.
14. K. B. Mullis et al. Target amplification by the polymerase chain reaction. *Proceedings of the National Academy of Sciences*, 83(8):2567–2571, 1986.
15. K. B. Mullis et al. Specific enzymatic amplification of DNA in vitro. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 51, pages 263–273. 1986.
16. Reinhard Heckel, Gediminas Mikutis, and Robert N. Grass. A characterization of the DNA data storage channel. *Sci. Rep.*, 9(9663), July 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45832-6.
17. Ammar Tareen and Justin B Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2020. doi: 10.1093/bioinformatics/btz921.

Supplementary Material

Supplementary Note SS1: Strand-aware alignment using anchor-based read filtering

Script S1. Code the strand-aware alignment using anchor-based read filtering for the assessment of oligo quality.

```
"""
### This code does:
##### Match forward reads to the left anchor,
##### Match reverse reads (revcomp) to the right anchor,
##### Combine them into one sequence logo plot to show varaibility
##### Usage example:
      python strand_aware_logo.py --fastq_dir fastq_pass --reference reference.fa --anchor_len 20 --output logo.png
### author: Robert Ramirez-Garcia
"""
import os
import argparse
from collections import Counter
from Bio import SeqIO
from Bio.Seq import Seq
import pandas as pd
import matplotlib.pyplot as plt
import logomaker
import matplotlib.patches as patches

def read_fasta(fasta_file):
   ### Reads a .fasta file and returns the sequence
   with open(fasta_file, 'r') as file:
      seq = []
      for line in file:
         if not line.startswith('>'):
            seq.append(line.strip())
   return ''.join(seq)

def get_sequences_from_fastq_folder(folder):
   ### Reads all sequences from all .fastq files within a folder
   seqs = []
   for fname in os.listdir(folder):
      if fname.endswith('.fastq') or fname.endswith('.fq'):
         for record in SeqIO.parse(os.path.join(folder, fname), "fastq"):
            seqs.append(str(record.seq))
   return seqs

def reverse_complement(seq):
   ### Return the reverse complement of a DNA sequence
   return str(Seq(seq).reverse_complement())

def anchor_extract(reference, sequence, anchor_len, mode):

   ### Extracts and aligns sequence to reference based on anchor mode.
   ### Returns a sequence of reference length, padded with '-' if needed.
   ### Modes: 'both', 'left', 'right'

   ref_len = len(reference)
   ref_left = reference[:anchor_len]
   ref_right = reference[-anchor_len:]

   if mode == 'both':
      if sequence.startswith(ref_left) and sequence.endswith(ref_right):
         mid_seq = sequence[anchor_len:len(sequence)-anchor_len]
         new_seq = ref_left + mid_seq + ref_right
         return new_seq[:ref_len].ljust(ref_len, '-')
   elif mode == 'left':
      if sequence.startswith(ref_left):
         mid_seq = sequence[anchor_len:]
         new_seq = ref_left + mid_seq
         return new_seq[:ref_len].ljust(ref_len, '-')
   elif mode == 'right':
      if sequence.endswith(ref_right):
         mid_seq = sequence[:len(sequence)-anchor_len]
         new_seq = mid_seq + ref_right
         return new_seq[-ref_len:].rjust(ref_len, '-')
   return None

def get_most_abundant_sequence(seq_list):
   if not seq_list:
      return ""
   return Counter(seq_list).most_common(1)[0][0]

def create_sequence_logo(seq_strings, reference, output_file="sequence_logo.png"):
   max_length = len(reference)
   alphabet = ['A', 'C', 'G', 'T', '-']
   counts_matrix = pd.DataFrame(0, index=range(max_length), columns=alphabet)
   for seq in seq_strings:
      for pos, char in enumerate(seq):
         if char in alphabet:
            counts_matrix.loc[pos, char] += 1
   freq_matrix = counts_matrix.div(counts_matrix.sum(axis=1), axis=0).fillna(0)
   consensus = ''
   for pos in range(max_length):
      max_freq = freq_matrix.iloc[pos].max()
      if max_freq > 0.5:
         consensus += freq_matrix.iloc[pos].idxmax()
      else:
         consensus += 'N'

   colour_code = {
      'A': '#008000', # Green
      'C': '#0000FF', # Blue
      'G': '#FFA500', # Orange
      'T': '#FF0000', # Red
      '-': '#808080', # Gray
      'N': '#808080'
   }

   most_abundant = get_most_abundant_sequence(seq_strings)

   fig, (ax_logo, ax_abundant, ax_ref, ax_cons) = plt.subplots(
      4, 1, figsize=(15, 10),
```

DRAFT

```
        gridspec_kw={'height_ratios': [4, 1, 1, 1]},
        sharex=True
    )

    logo = logomaker.Logo(freq_matrix, ax=ax_logo, color_scheme=colour_code)
    logo.style_spines(visible=False)
    ax_logo.set_ylabel('Frequency')
    ax_logo.set_title('Sequence Logo')

    for pos, char in enumerate(most_abundant):
        if char in colour_code:
            colour = colour_code[char]
            rect = patches.Rectangle((pos-0.5, 0), 1, 1, facecolor=colour, alpha=0.6)
            ax_abundant.add_patch(rect)
            ax_abundant.text(pos, 0.5, char, ha='center', va='center', fontsize=10, fontweight='bold', color='white')
    ax_abundant.set_ylim(0, 1)
    ax_abundant.set_yticks([])
    ax_abundant.set_title('Most Abundant Sequence')
    ax_abundant.set_frame_on(False)

    for pos, char in enumerate(reference.ljust(max_length, '-')):
        if char in colour_code:
            colour = colour_code[char]
            rect = patches.Rectangle((pos-0.5, 0), 1, 1, facecolor=colour, alpha=0.6)
            ax_ref.add_patch(rect)
            ax_ref.text(pos, 0.5, char, ha='center', va='center', fontsize=10, fontweight='bold', color='white')
    ax_ref.set_ylim(0, 1)
    ax_ref.set_yticks([])
    ax_ref.set_title('Reference Sequence')
    ax_ref.set_frame_on(False)

    for pos, char in enumerate(consensus):
        if char in colour_code:
            colour = colour_code[char]
            rect = patches.Rectangle((pos-0.5, 0), 1, 1, facecolor=colour, alpha=0.6)
            ax_cons.add_patch(rect)
            ax_cons.text(pos, 0.5, char, ha='center', va='center', fontsize=10, fontweight='bold', color='white')
    ax_cons.set_ylim(0, 1)
    ax_cons.set_yticks([])
    ax_cons.set_xlabel('Position')
    ax_cons.set_title('Consensus Sequence')
    ax_cons.set_frame_on(False)

    plt.tight_layout()
    plt.savefig(output_file, dpi=300)
    print(f"Most abundant sequence: {most_abundant}")
    print(f"Consensus sequence: {consensus}")
    print(f"Sequence logo saved as {output_file}")

if __name__ == "__main__":
    parser = argparse.ArgumentParser(description="Strand-aware anchor alignment and sequence logo generator.")
    parser.add_argument("--fastq_dir", type=str, default="fastq_pass", help="Directory with .fastq files")
    parser.add_argument("--reference", type=str, required=True, help=".fasta / .fa file with reference sequence")
    parser.add_argument("--anchor_len", type=int, default=20, help="Length of anchor region (nt)")
    parser.add_argument("--output", type=str, default="sequence_logo.png", help="Output image filename")
    args = parser.parse_args()

    reference = read_fasta(args.reference)
    all_seqs = get_sequences_from_fastq_folder(args.fastq_dir)
    print(f"Total sequences read: {len(all_seqs)}")

    selected = []

    for seq in all_seqs:
        aligned = anchor_extract(reference, seq, args.anchor_len, mode="left")
        if aligned:
            selected.append(aligned)
            continue

        rev_seq = reverse_complement(seq)
        aligned_rc = anchor_extract(reference, rev_seq, args.anchor_len, mode="right")
        if aligned_rc:
            selected.append(aligned_rc)

    print(f"Sequences passing anchor filter: {len(selected)}")
    if not selected:
        print("No sequences passed the anchor filter.")
        exit(1)

    ### Precomputes most abundant and consensus for clean reporting
    most_abundant = get_most_abundant_sequence(selected)

    max_length = len(reference)
    counts_matrix = pd.DataFrame(0, index=range(max_length), columns=['A', 'C', 'G', 'T', '-'])
    for seq in selected:
        for pos, char in enumerate(seq):
            if char in counts_matrix.columns:
                counts_matrix.loc[pos, char] += 1
    freq_matrix = counts_matrix.div(counts_matrix.sum(axis=1), axis=0).fillna(0)

    consensus = ''
    for pos in range(max_length):
        max_freq = freq_matrix.iloc[pos].max()
        if max_freq > 0.5:
            consensus += freq_matrix.iloc[pos].idxmax()
        else:
            consensus += 'N'

    ### Outputs all expected logs
    print(f"Most abundant sequence: {most_abundant}")
    print(f"Reference sequence used in plot: {reference}")
    print(f"Consensus sequence: {consensus}")

    ### Creates and saves the logo
    create_sequence_logo(selected, reference, args.output)
```