Nucleic Acids Research, 2026, **54**, gkag433
https://doi.org/10.1093/nar/gkag433
Critical Reviews and Perspectives

Cross-contamination of commercial oligonucleotides with library-structured sequences

Manouk Gerritsen [1,2,3], Partho Sarothi Ray [1,2,4,5], Thileepan Sekaran [1], Thomas Schwarzl [1], Matthias W. Hentze [1,2,*], Andreas E. Kulozik [1,2,4,*]

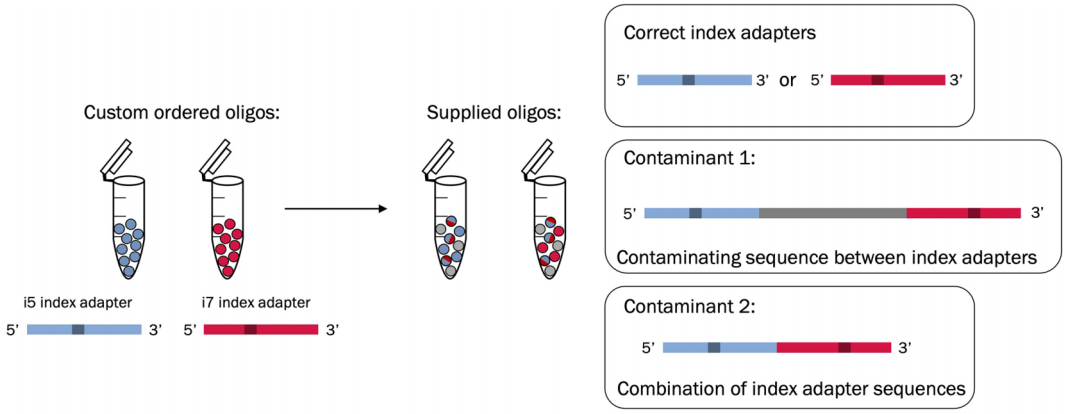
[1]European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2]Molecular Medicine Partnership Unit (MMPU), Im Neuenheimer Feld 350, 69120 Heidelberg, Germany
[3]Collaboration for joint PhD degree between EMBL and Faculty of Biosciences, Heidelberg University, 69120 Heidelberg, Germany
[4]Department of Paediatric Oncology, Haematology and Immunology, Hopp Children's Cancer Center (KiTZ), Heidelberg University & Clinical Cooperatiuon Unit "Pediatric Leukemia, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 430, 69120 Heidelberg, Germany
[5]Present address: SABNP Laboratory, INSERM U1204, Université d'Evry-Paris-Saclay, Evry 91025, France

*To whom correspondence should be addressed. Email: hentze@embl.org
Correspondence may also be addressed to Andreas E. Kulozik. Email: andreas.kulozik@med.uni-heidelberg.de

Abstract

Custom oligonucleotides (oligos) play a crucial role in molecular biology applications, including sequencing, polymerase chain reaction (PCR), and library construction. For highly sensitive techniques such as next-generation sequencing (NGS) library preparation, oligos must exhibit exceptional purity and accuracy. Here, we provide a perspective on the presence of unrelated, contaminating sequences in custom-made oligos using a case study of an RNA–protein crosslinked immunoprecipitation (CLIP-seq) experiment. These contaminants led to artefactual results in NGS data, rendering their results uninterpretable. Systematic investigation identified a contaminating human immunodeficiency virus (HIV) *gag* sequence embedded between Illumina-compatible adapter sequences, forming a functional 'library-structured' molecule capable of amplification and sequencing. Additionally, other nonspecific sequences, including fragments from unrelated oligos, were detected. By sharing our analysis strategy and outcomes, we aim to raise awareness within the research community and emphasize the need for more stringent quality controls in oligo synthesis and validation by commercial suppliers, safeguarding academic experimenters' time and resources. Furthermore, this perspective provides a systematic approach for the identification of the source of contaminating sequences in custom-made oligos designed for NGS library preparations.

Graphical abstract



Introduction

Oligonucleotides in research and therapy

Oligonucleotides (oligos) are short single- or double-stranded DNA or RNA molecules, which are of vital importance in various research and clinical applications. They are usually produced in the laboratory through solid-phase synthesis and can be synthesized as single-stranded molecules with precise sequences [1]. Double-stranded oligos are obtained by annealing

Received: October 31, 2025. **Revised:** March 28, 2026. **Accepted:** March 31, 2026

two complementary strands after synthesis, cleavage and deprotection, and purification. Oligos serve as versatile tools in various applications in research, genetic testing, forensics, and therapeutics. They are used extensively in applications such as polymerase chain reaction (PCR), DNA and RNA sequencing, molecular cloning, artificial gene construction, molecular probing, genome editing, and gene knockdown [2]. Increasingly, they are being tried for clinical applications as antisense oligonucleotides (ASOs), small interfering RNAs (siRNAs), aptamers, and guides for genome editing [3].

Oligos possess the unique capability to hybridizing specifically to their complementary molecules such as DNA or RNA, leading to the formation of duplexes, or in some cases, higher-order hybrids. This characteristic allows the use of oligos as probes for identifying specific DNA or RNA sequences as well as amplifying them. Moreover, oligos can be joined to DNA or RNA sequences by the formation of 5′–3′ phosphodiester bonds, utilizing a variety of ligase enzymes, which has increased their versatility. Current techniques for utilizing oligos include next-generation sequencing (NGS), fluorescent *in situ* hybridization (FISH), synthesis of artificial genes, and clustered regularly interspaced short palindromic repeats (CRISPR)-based genome editing (Fig. 1).

Application of oligonucleotides in CLIP-seq

NGS-based techniques have revolutionized genomics, providing deep insights into genome structure, function, and expression. Specialized applications of NGS have allowed exploration of more and more specific and challenging questions in biology, such as transcriptome dynamics and interactomics. For example, RNA–protein crosslinking and immunoprecipitation (CLIP) technologies have been developed to identify the binding sites of RNA-binding proteins (RBPs) on their RNA targets in cells using an NGS-based approach (CLIP-seq) [4]. Advances in CLIP-seq methodology over the last 20 years since the original study [5] have addressed technical challenges and improved the resolution, sensitivity, specificity, and convenience of the method [4].

To amplify the RNA fragments that are bound by their cognate RBPs in CLIP-seq and prepare cDNA libraries, different RNA and DNA adapter oligonucleotides (oligos) and primers are used throughout the protocol. The correct and efficient addition of each of these sequences to the RNA or cDNA and the complementarity of these sequences to the oligos used in the subsequent steps is critical for obtaining reproducible results. The reliable outcome of CLIP experiments is thus dependent on the specificity and purity of all these oligos.

In order to understand the crucial usage of oligos in the different steps of CLIP-seq, it is important to have an overview of the CLIP-seq workflow. Essentially, in most CLIP-seq variants similar steps are involved to purify RBP–RNA complexes and subsequently sequence the RNA fragments to profile the sites of intermolecular contact on a transcriptome-wide level [6, 7]. The most widely used CLIP variants enable RBP binding sites to be determined with single-nucleotide resolution. To this end, RBPs are first covalently crosslinked to their RNA targets by controlled ultraviolet (UV) irradiation of cells (Fig. 2B). The crosslinked RNAs are then partially fragmented by limited RNase treatment. After immunoprecipitation using specific antibodies against the bound RBP, adapter ligation to the RNA and sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS–PAGE), the RBP-associated RNA fragments are reverse transcribed into cDNA (Fig. 2C–G). Then, sequencing adapter oligos are ligated to the cDNA to enable PCR amplification and the generation of cDNA libraries that are subjected to high-throughput NGS sequencing (Fig. 2H–J).

As the CLIP-seq workflow is dependent on the usage of multiple adapter and primer oligos, the purity, integrity, and specificity of oligos is imperative for the success of a CLIP-seq experiment. For example, it is known that cross-contamination of index adapters can lead to the misassignment of reads when samples are multiplexed, through index cross-talk or index hopping [8, 9]. Similarly, the presence of failed synthesis products, including truncated oligos or other contaminants, can potentially interfere with the amplification of the desired sequence. Although commercially synthesized research-grade oligos are marketed with several levels of quality control, including mass spectrometry, and multiple purification options such as high-performance liquid chromatography (HPLC) or PAGE, such oligos have been reported to be frequently cross-contaminated with CRISPR guide sequences and other unrelated oligos [10].

Contamination of commercially supplied oligonucleotides: a case study

Here, we provide a perspective about the criticality of the purity and integrity of oligonucleotides to an NGS-based technique such as CLIP-seq. We use a case study of a complex contamination of commercially synthesized sequencing adapter oligos from a reputed supplier used for library amplification in CLIP experiments to illustrate the above perspective. The contamination was identified during data analysis and required a systematic investigation to trace its origin to the supplied oligos. The adapter oligos used were reported to have passed all quality controls advertised by the supplier, yet contained various erroneous sequences, one of which caused a significant over representation of artefactual sequences in the CLIP data. While we report a CLIP-seq experiment as a use case to illustrate the pitfalls of the presence of contaminating oligos in an experimental workflow utilizing multiple sets of oligos, such contaminations are of critical relevance for many NGS-based experiments involving commercially supplied oligos.

While it is unknown how frequently such contaminations occur in custom oligonucleotides, reports on how to systematically troubleshoot such issues are lacking. Therefore, this case study highlights the need for the scientific community to remain vigilant about potential failures in the quality control of commercially provided oligos, which can lead to confounding results and unnecessary expenditure of time and resources in NGS-based experiments.

An abundant contaminant sequence identified in seCLIP data

The case study is based on single-end enhanced CLIP experiments (seCLIP) aimed at identifying RNA targets of two RBPs in human hepatocellular carcinoma (HuH7) cells. The seCLIP experiment was performed according to a published protocol with modifications (Fig. 2; see Supplemental data for details of the methods and Supplementary Table S1) [11]. Briefly, this method relies on immunoprecipitation of the RBP of interest followed by preparation of a sequencing library from the crosslinked RNAs. The crosslink-induced termination of cDNA synthesis and adapter ligation strategy ensures that the crosslink site is positioned at the beginning of the first

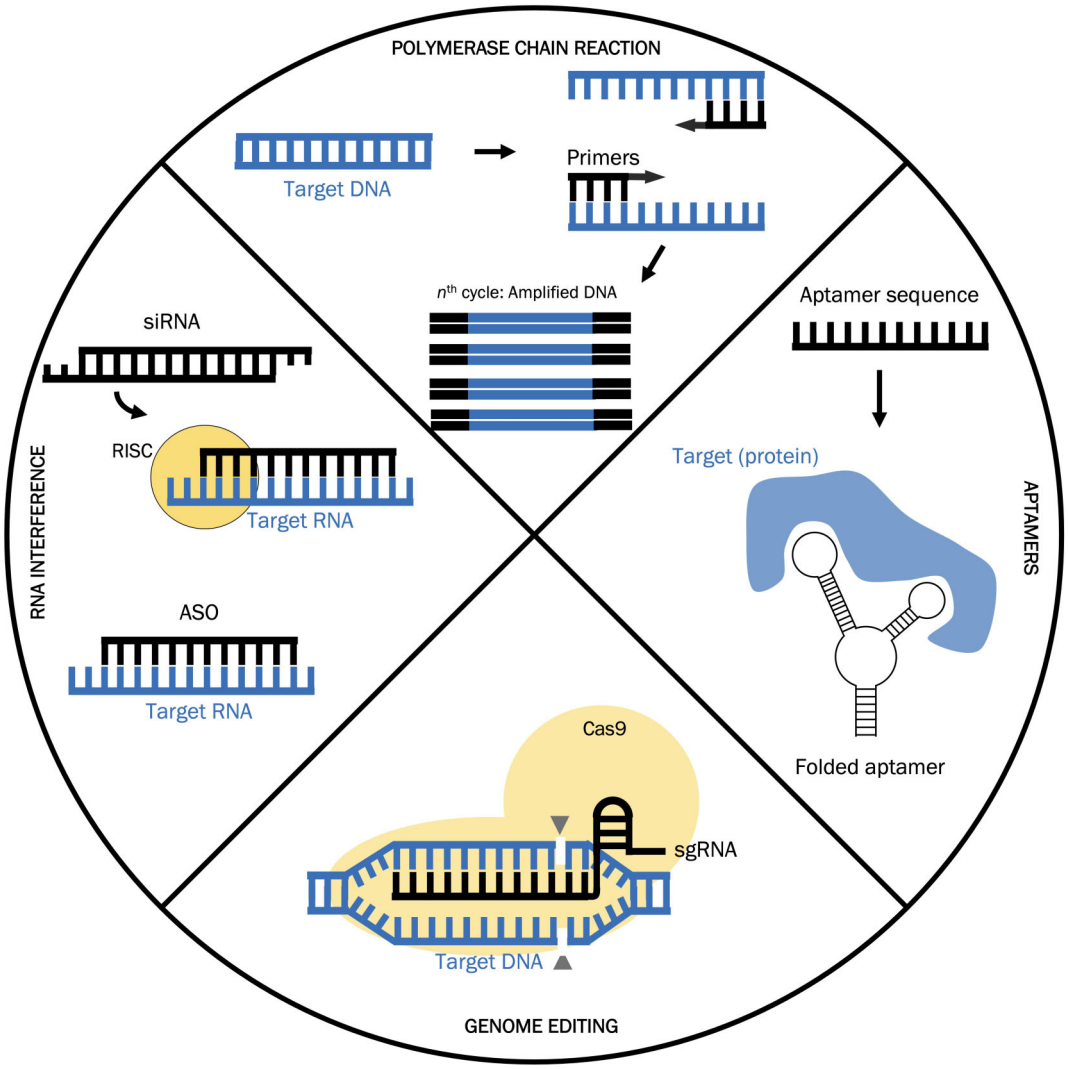


Figure 1. Research and clinical applications of oligonucleotides. Examples of application of oligonucleotides in the laboratory and clinic, including in polymerase chain reaction (widely used in molecular cloning, NGS, and diagnostics), as aptamers for studying molecular interactions, as guide RNAs in genome editing, and as siRNAs and antisense oligonucleotides (ASOs) for gene knockdown.

sequencing read. A unique molecular identifier (UMI), consisting of 10 random nucleotides, is part of the adapter oligo that is ligated to these truncated cDNAs to uniquely tag each molecule, which places the UMI upstream of the insert. This information is thus captured in the same sequencing read as the crosslink site. During library preparation experimental indexes (or barcodes) of eight nucleotides are added to the cDNAs via the primer oligos to uniquely label each sample, which enables different libraries to be multiplexed and sequenced in the same run. Finally, these data are assembled and analysed using computational tools to identify the RNA target and specific crosslinking site(s). The inclusion of UMIs within the cDNA adapter oligos also allows for the quantification of unique cDNAs in the library and removal of PCR duplicates. Here, a combinatorial dual indexing strategy to label the libraries was adopted, using four different Illumina-compatible forward (i5) and reverse (i7) index adapters (hereafter also abbreviated as D50_ and D70_ in figures, respectively) (Fig. 2A) for the library amplification step, permitting the multiplexing of the samples. Each double stranded molecule in the sequencing library therefore had a 'library-structure' consisting of a cDNA of interest, flanked by a 5′ UMI sequence and sequencing compatible i5 and i7 adapter sequences to the 5′ and 3′ of the cDNA sequence (Fig. 2J).

Using this method, cDNA libraries for RNAs binding to two different RBPs, SERBP1 and PCBP1, were prepared together with appropriate controls. The experimental pipeline was quality controlled at multiple levels. In particular, immunoprecipitation of the RBP with IP-grade antibodies was visualized by western blot. The position of the RBP on the western blot was used to determine the region of the membrane (RBP + 75 kDa) to be recovered, to capture and enrich for the desired RBP–RNA complexes. The number of PCR cycles required for library preparation was also optimized in test PCRs to avoid overamplification. The cDNAs in the final libraries were size selected and purified as per the protocol before sequencing.

For the analysis of the sequencing data, the libraries were first demultiplexed using the indexes and the adapter sequences were removed from the reads (Fig. 2K). As the organization of the sequence read is predefined (see Fig. 2J), the UMIs were then extracted by taking the first 10 nucleotides

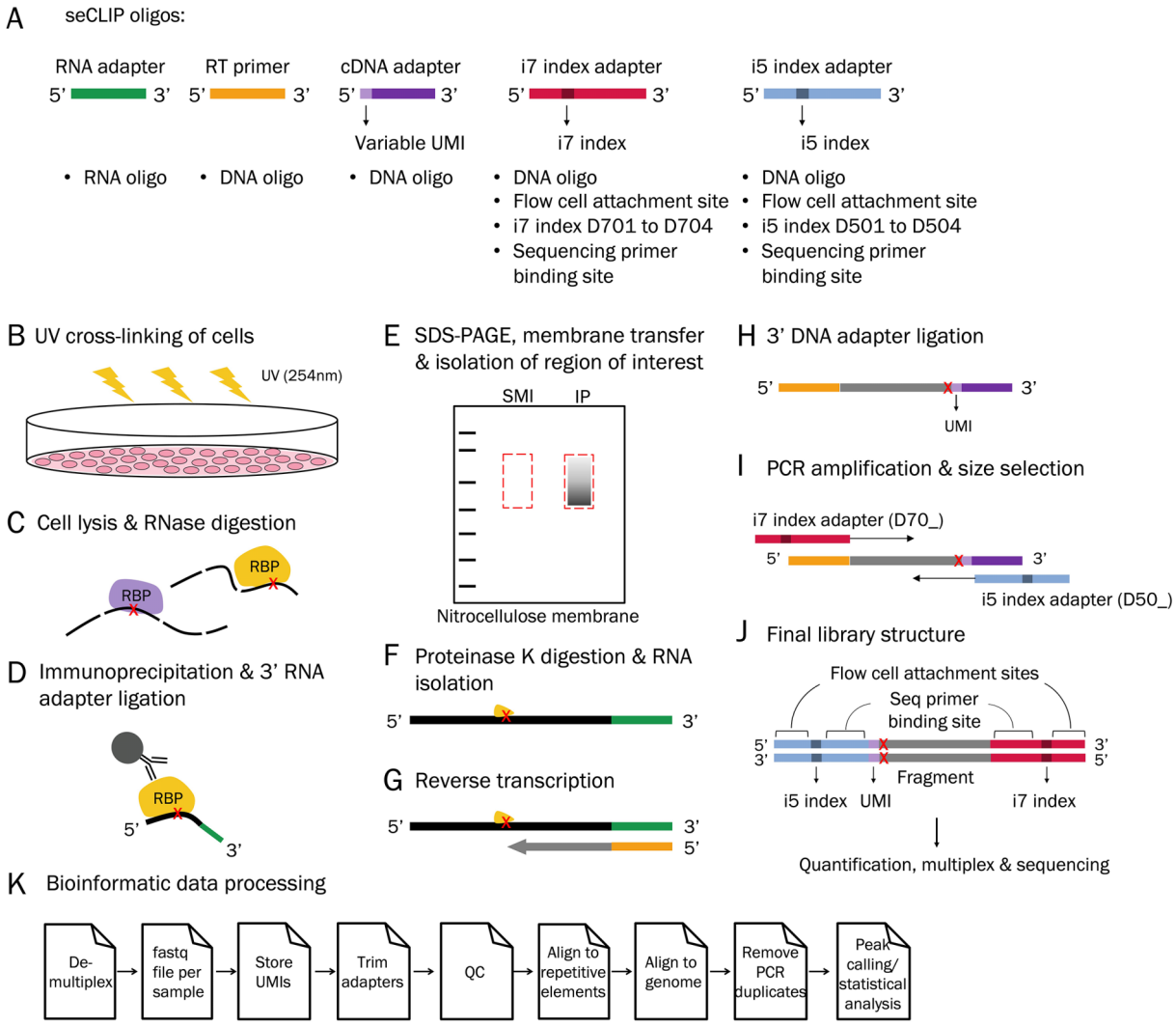


Figure 2. Schematic of the seCLIP experimental protocol and bioinformatic analysis pipeline. (**A**) Schematic representation of the different oligos (one oligoribonucleotide and four types of oligodeoxyribonucleotides) used in these CLIP protocol. Important characteristics/features are annotated and listed. (**B–J**) Overview of the seCLIP experimental protocol [11]. (**B**) Live cells are irradiated with UV light at 254 nm to induce covalent crosslinking between RNA and proteins that are in direct contact. (**C**) Following cell lysis, the RNA is partially digested into fragments by RNAse I endoribonuclease. (**D**) The RBP of interest is immunoprecipitated using antibody-coupled magnetic beads and after several washes and a 3′-dephosphorylation step an RNA adapter oligo (green) is ligated to the 3′ end of co-immunoprecipitated RNA fragments in preparation for reverse transcription. (**E**) RBP–RNA complexes are separated by SDS–PAGE and transferred to a nitrocellulose membrane. A region of the membrane up to 75 kDa above the RBP size is excised to capture the RNAs bound by the RBP. SMI: size-matched input control. (**F**) Proteinase K digestion removes the crosslinked RBP and the RNA is extracted from the membrane. (**G**) Reverse transcription to cDNA is performed with a primer (orange) complementary to the RNA adapter ligated instep D. Truncated cDNAs are generated due to the peptide at the crosslink site (red X). (**H**) Ligation of 3′ DNA adapter (purple) to the cDNA molecules. This cDNA adapter contains a 10-nucleotide random-mer at the 5′ end, which is the UMI (indicated in light purple). (**I**) Adapter-ligated cDNA libraries are PCR amplified using DNA oligos (blue and red) that comprise Illumina platform-compatible adapter sequences along with an in-line experimental index (darker shade) and sequences complementary to either the RT primer or cDNA adapter. Amplified cDNA libraries are subsequently size selected and purified to select cDNAs of 175–350 bp and ensure complete removal of adapter ligation products. (**J**) Final structure of the seCLIP library fragments including adapters with an i5 and i7 index flanking either side of the fragment of interest. The libraries are quantified, multiplexed, and subjected to high-throughput sequencing. (**K**) Representative example of an seCLIP data analysis workflow. Briefly, reads are demultiplexed, the UMIs are extracted from the fastq files, and the adapter sequences are trimmed. After quality control (QC), reads are typically first aligned to a database of repetitive elements to remove these sequences and only the unmapped reads are mapped to the genome of interest. PCR duplicates are removed using UMI information and finally the reads are used as input for statistical analysis and calling of interacting RNAs and binding sites.

of the read, also used later to determine PCR duplicates to be collapsed to a single read. After alignment to Repbase [12], a database of repetitive eukaryotic DNA sequences, to remove repeat elements from the sequences, the reads were mapped to the human genome. Unexpectedly, data analysis revealed high PCR duplication rates across all samples. Genome alignment demonstrated that a substantial number of reads did not map to the human reference genome (Fig. 3A). The proportion of unmapped reads ranged from ~25% to 90% across different samples. Further analyses showed that the most overrepresented unmapped sequence was identical in all seCLIP libraries and revealed 100% identity to nucleotide positions 1703–1737 of the human immunodeficiency virus type 1 (HIV-1) (Fig. 3B), corresponding to the group specific antigen (*gag*) gene (Fig. 3C). The *gag* gene encodes the gag polyprotein that is cleaved into the core structural compo-

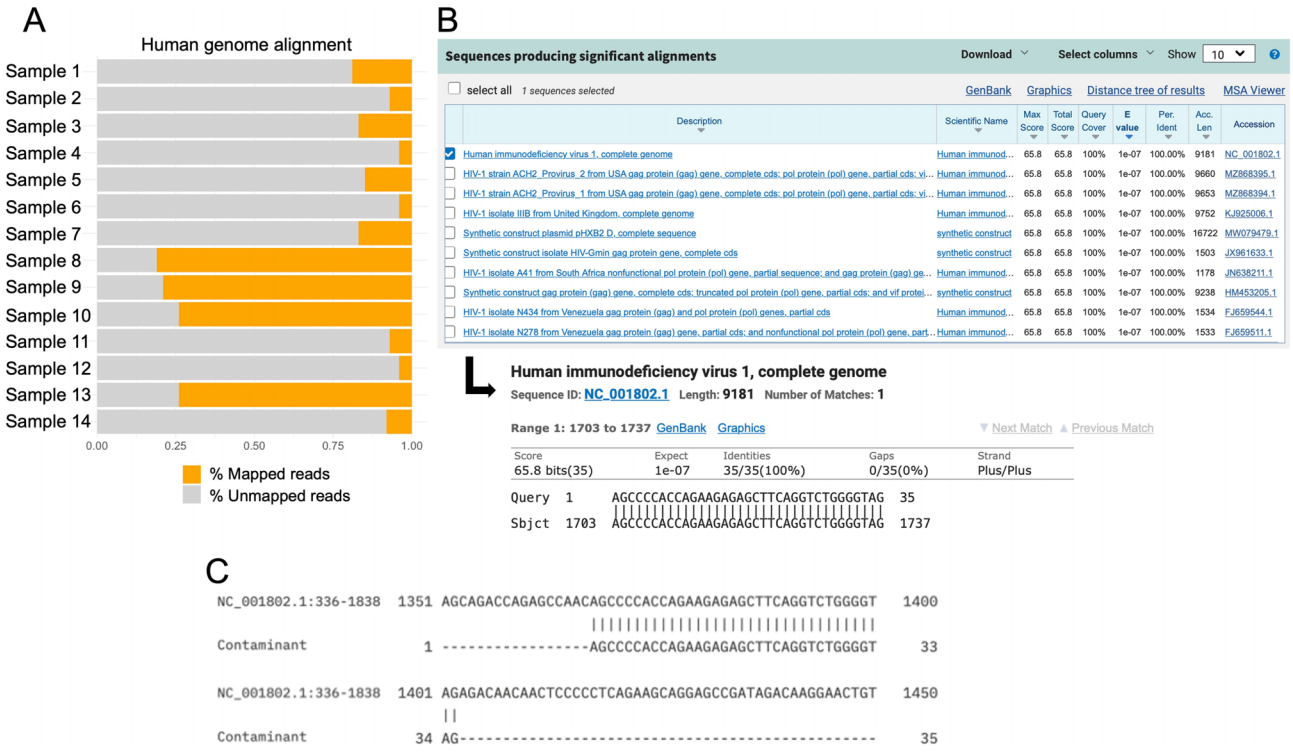


Figure 3. Identification of over-represented HIV *gag* sequences in seCLIP reads of human (HuH7) cells. (**A**) Stacked bar graph showing the percentage of reads in the 14 seCLIP samples that were mapped (orange) or unmapped (grey) to the human genome (GRCh38). (**B**) NCBI Nucleotide BLAST (blastn) results where the most over-represented unaligned seCLIP read was queried against the default nr/nt database (top). Detailed pairwise alignment with HIV-1 (NC_001802.1) is shown in the bottom panel. (**C**) EMBOSS Needle pairwise alignment of the contaminating seCLIP read to the *gag* gene of HIV-1 (NC_001802.1, nucleotide positions 336-1838).

nents of HIV-1 (and all other retroviruses) and is required and sufficient to drive virion assembly [13]. As this overrepresentation of a single, completely unrelated sequence was unexpected, we concluded that this result was artefactual. While it is easy to identify the presence of such an artefact, it is challenging to identify its source, considering the multiple inputs into the experimental pipeline. Therefore, a methodology had to be developed to systematically exclude multiple possibilities and finally identify the source of this likely artefactual result.

The HIV *gag* contaminant sequence lacks a UMI

As a first step, we explored the possibility that the HIV *gag* sequence was already present in the HuH7 cells, as retroviral genomic sequences are sometimes integrated into cellular genomes. Therefore, we performed reverse transcription – polymerase chain reaction (RT-PCR) with primers designed against the HIV-1 *gag* gene on RNA isolated from the same HuH7 cell lysates that were used for the seCLIP experiments. While *β-actin* mRNA used as positive control was readily amplified from all samples, the HIV *gag*-specific primers failed to amplify a product (Fig. 4A), indicating that the contaminating sequence was not present in the HuH7 cellular RNA and had not been introduced during cell culture. Consistent with this result, the contaminating sequence identified in these CLIP data corresponded to only one particular fragment of the HIV *gag* gene (Fig. 3C) and no other sequences from the HIV-1 *gag* or other parts of the HIV-1 genome were found, indicating the absence of functional HIV-1 (*gag*) RNA. Together, these findings ruled out HIV as the contaminating source. Also, the presence of a single, short sequence from the HIV *gag* gene was inconsistent with the possibility of introduction of HIV genomic RNA from any reagent or material during the CLIP process as a previous report describing the presence of contaminating *Acinetobacter johnsonii XBB1* (CP010350.1) sequences in CLIP data, originating from contaminated transfer membrane, had found sequences covering the entire *A. johnsonii* genome [14].

We next excluded the possibility that the contamination originated from the sequencing facility. The remaining aliquots of the seCLIP libraries that had been submitted to the sequencing facility were subjected to PCR reactions with a reverse primer specifically designed to identify the contaminating *gag* sequence and a primer against the 5′ sequence common to all forward (i5) index adapters (we refer to this as the i5 flow cell primer) (see Figs 2J and 4B, top). A PCR product of ∼100 bp was amplified from all tested samples (Fig. 4B). This length corresponded to the expected size of the contaminating HIV sequence and Sanger sequencing of this PCR product following gel-purification confirmed it as the contaminant HIV *gag* DNA sequence (Fig. 4B). Identical results were obtained over multiple experiments where a *gag* sequence primer was used in combination with an i7 flow cell primer or with full length i5 or i7 index adapters (data not shown). These analyses therefore demonstrated that the contaminating sequence was introduced during library preparation and before submission of the libraries to the sequencing facility.

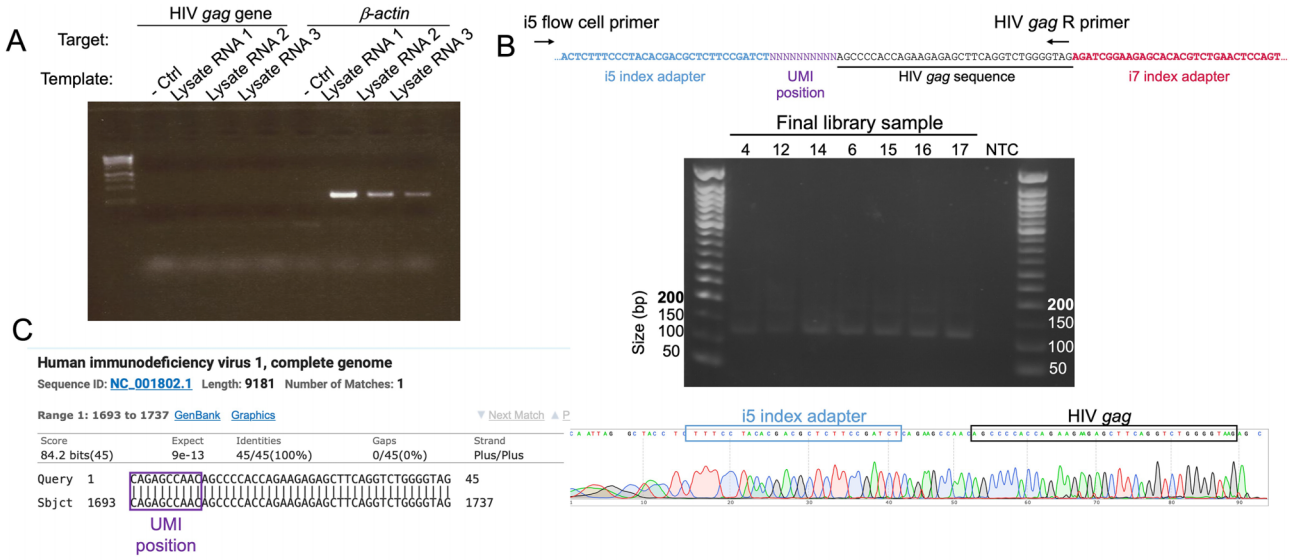


Figure 4. The contaminant HIV *gag* sequence was introduced post cell culture and lacks a UMI. (**A**) Gel electrophoresis of RT-PCRs of RNA isolated from seCLIP HuH7 cell lysates using primers designed against a region in the HIV *gag* gene (left) or β-actin as positive control (right). (**B**) Top: The contaminant seCLIP read from our experiment is shown with the HIV *gag* sequence flanked by index adapters.The expected position of the UMI is displayed in purple as 10 N's to reflect the 10 random nucleotides. The locations of primers used in the PCR reaction shown below are indicated. Note, the sequences of the i5 and i7 index adapters are only partially shown. Middle: Gel electrophoresis of PCRs performed on seCLIP final library samples amplified with an HIV *gag* specific primer and a primer against the flow cell binding sequence of i5 index adapters. Bottom: Sanger sequencing result shown as chromatogram data with the contaminating HIV *gag* sequence and (partial) i5 index adapter sequences indicated. (**C**) Detailed pairwise alignment of nucleotide BLAST (blastn) results of B showing 100% sequence identity to HIV-1. The 10 nt sequence in the UMI position, but found to be invariably aligned to HIV-1, is highlighted with a box.

Importantly, the raw seCLIP reads as well as the Sanger sequencing results demonstrated absence of any UMI sequence between the i5 index adapter sequence and the HIV *gag* sequence in these reads (Fig. 4C). Instead of the UMI sequence there was a 10 bp sequence from the HIV *gag* gene in the UMI position, giving rise to a continuous 45 bp HIV *gag* sequence (Fig. 4C). As the UMIs are used for identification of PCR-duplicated reads, the absence of a UMI explained the observed high duplication rate of the libraries with the contaminating sequence. More importantly, this finding indicated that it was unlikely that the contamination was present in the steps prior to cDNA adapter oligo ligation, because all RNA fragments obtained from earlier steps, including any potential nonspecific DNA or RNA contaminants, would otherwise have been ligated to a UMI-containing adapter oligo after reverse transcription to cDNA (Fig. 2H).

Despite the absence of a UMI, the contaminating sequence had the correct library-structure, i.e. a cDNA sequence flanked by index adapter sequences (Figs 2J and 4B), which is strictly necessary for it to bind to the flow cell and be sequenced on the NGS platform. The presence of the erroneous HIV *gag* sequence between the index adapters resulted in a DNA fragment of 180 bp (45 bp HIV *gag* + 70 bp i5 index adapter + 65 bp i7 index adapter), which is within the 175–350 bp size range of cDNA library fragments selected for sequencing.

The index adapter oligos contain the contaminating HIV *gag* sequence

Collectively, our findings suggested that the contamination was likely introduced at one of the steps in the CLIP protocol subsequent to the cDNA synthesis (i.e. Fig. 2H–I). To this end, we first questioned whether the cDNA adapter oligo (see Fig. 2H) could have been the source of the HIV *gag* sequence. We used the cDNA adapter as template in PCR reactions with two different combinations of i5 and i7 index adapters (indicated as D501-4 and D701-4 in Fig. 5). The cDNA adapter sequence should only have had a binding site for the i5 but not i7 index adapter and PCRs should therefore not generate a product.

By contrast, we found that a product of about 100 bp was amplified together with an unexpected ∼100–150 bp product that also appeared in the no template control (NTC) reactions (Fig. 5A). Different water sources were tested, but all showed this nonspecific amplification, suggesting that this result was not caused by the use of contaminated water in the PCR reactions. Other PCR components such as dNTPs, buffers, and the DNA polymerase were also replaced to exclude the possibility of contamination of these reagents.

In subsequent PCR reactions, we used all i5 and i7 index adapter oligos that had been used in the seCLIP experiments as primers together with an HIV *gag* contaminant specific primer (Fig. 5B) to test whether the index adapter oligos could have been the source of the contamination. Amplification from these reactions would indicate that the index adapter oligos were both acting as primers and providing the template for these PCR reactions. Remarkably, here we found that five out of eight reactions generated a product (Fig. 5B, top), indicating that the index adapters were the source of the template for these amplicons. Sanger sequencing of the gel-extracted ∼100 bp PCR product revealed that this fragment contained the HIV *gag* sequence that contaminated the seCLIP data and was directly flanked by the sequence of the index adapter (Fig. 5B, bottom). These sequences matched with the HIV *gag* sequence and i7 index adapter as shown in Fig. 4.

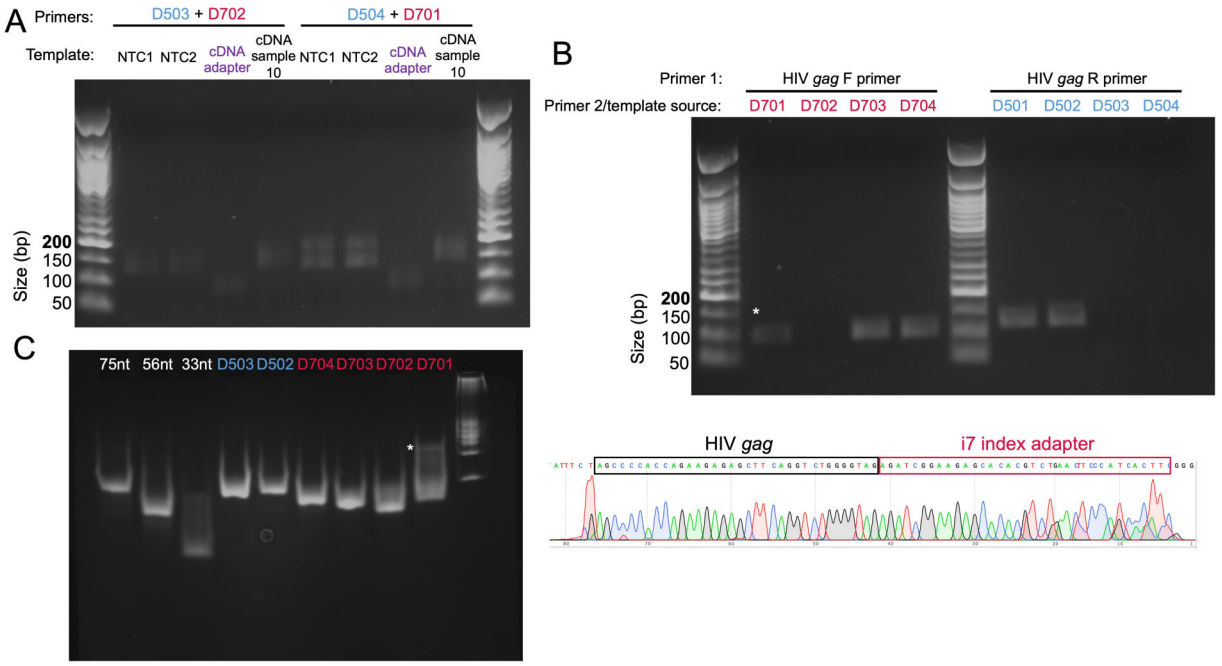


Figure 5. The HIV *gag* sequence is part of index adapter oligo molecules. (**A**) Gel electrophoresis of PCR reactions with two index adapters (forward i5 in blue indicated as D503/4 and reverse i7 in red as D701/2) performed on cDNA adapter, two different NTCs and cDNA library (positive control). (**B**) Top: Agarose gel electrophoresis of PCRs performed with an HIV *gag* contamination primer and i5 or i7 index adapter. The index adapters function as both primer and template in these reactions. Bottom: Sanger sequencing result of the PCR product generated with an HIV *gag* primer and i7 index adapter D701 gel purified from lane 2 (highlighted with asterisk). Result is shown as chromatogram data with the contaminating HIV *gag* sequence and (partial) i7 index adapter sequences indicated. (**C**) 5 μg of the indicated oligos were run on a PAGE gel. Lanes marked with 75 nt, 56 nt, and 33 nt contain 5 μg unrelated oligos of indicated length from a different supplier as size markers. The additional band present in the i7 index adapter D701 is highlighted with an asterisk.

These data thus conclusively demonstrated that the contaminating HIV *gag* sequence was part of the index adapter oligo. The different levels of product amplification with different index adapter oligos suggest that the contaminating sequence is present at different levels in different supplied oligos.

The HIV *gag* contamination was found in all samples of the seCLIP experiment to varying degrees (Fig. 3A). Because demultiplexing of the data relies on the unique combination of indexes for each sample in the experiment (i.e. the combinatorial dual indexing approach), at least one oligo in each adapter duo was contaminated with this sequence. Indeed, we found that the HIV *gag* contaminant sequence is present in at least five out of eight supplied oligos (Fig. 5B), which can directly explain 93% of the seCLIP data (13 out of 14 samples). Although we could not exactly quantify the level of contamination, the high proportion of unmapped reads in the different libraries (Fig. 3) indicates that this proportion was substantial.

We next compared the length of the delivered, contaminated oligos with the expected length of the index adapters by PAGE, using oligos of known length as size markers (Fig. 5C). The ordered sequences of the i5 and i7 index adapters were 70 nt and 65 nt in length, respectively. In contrast, the delivered PAGE-purified oligos all ran at slightly different positions thus indicating heterogeneity in length and sequence. Strikingly, in one of the oligos (Fig. 5C, lane 9 with i7 index adapter D701) we could even observe an additional band indicating the presence of another oligo or contaminant. Whereas smaller oligo by-products can occur as a result of failed oligo synthesis, the presence of larger products is unexpected, especially in PAGE-purified oligos that have been validated by the supplier to correspond to the correct mass as assessed by mass spectrometry (Supplementary data).

HIV *gag*, as a library-structured molecule, is present in index adapter oligos

The full adapter sequences must be present on both sides of the cDNA fragment of interest in order to attach to the sequencing flow cell and be sequenced on the NGS platform as these contain the sequences necessary for flow cell binding and the sequencing primer annealing site (Fig. 2J). Therefore, for inclusion in the seCLIP data, the contaminant sequence must have had this organization. In support of this, our results so far identified that the HIV *gag* contaminant sequence was part of the index adapter molecule (example shown for D701 in Fig. 5B). To test this hypothesis, we set out to investigate whether this contaminant sequence was already present in the supplied oligos as a 'library-structured' sequence, in which the contaminant is present between two index adapters (see also Fig. 2J). Because these adapters were used for seCLIP library amplifications, the contaminant would thus have been introduced into all of the sequences. In an unbiased approach to target away from the actual contaminant sequence, we used the i5 and i7 flow cell primers as detecting agents in PCR experiments where a single index adapter was used as a template (Fig. 6A). Unless the annealing site for the second primer is present too, this should not generate any PCR product. If the HIV *gag* and other index adapters sequence were both part of the same molecule the expected product would be 180 bp. Indeed, in all reactions where one of four index adapters was used as template, we amplified a product but not in the NTC PCR. Sequencing of the gel-purified PCR products (Fig. 6C–E) identified the 45 bp contaminant HIV *gag* sequence and part of both index adapters in the ~200 bp product from the reaction where i5 index adapter D501 was used as template (Fig. 6C). Thus, even without using HIV

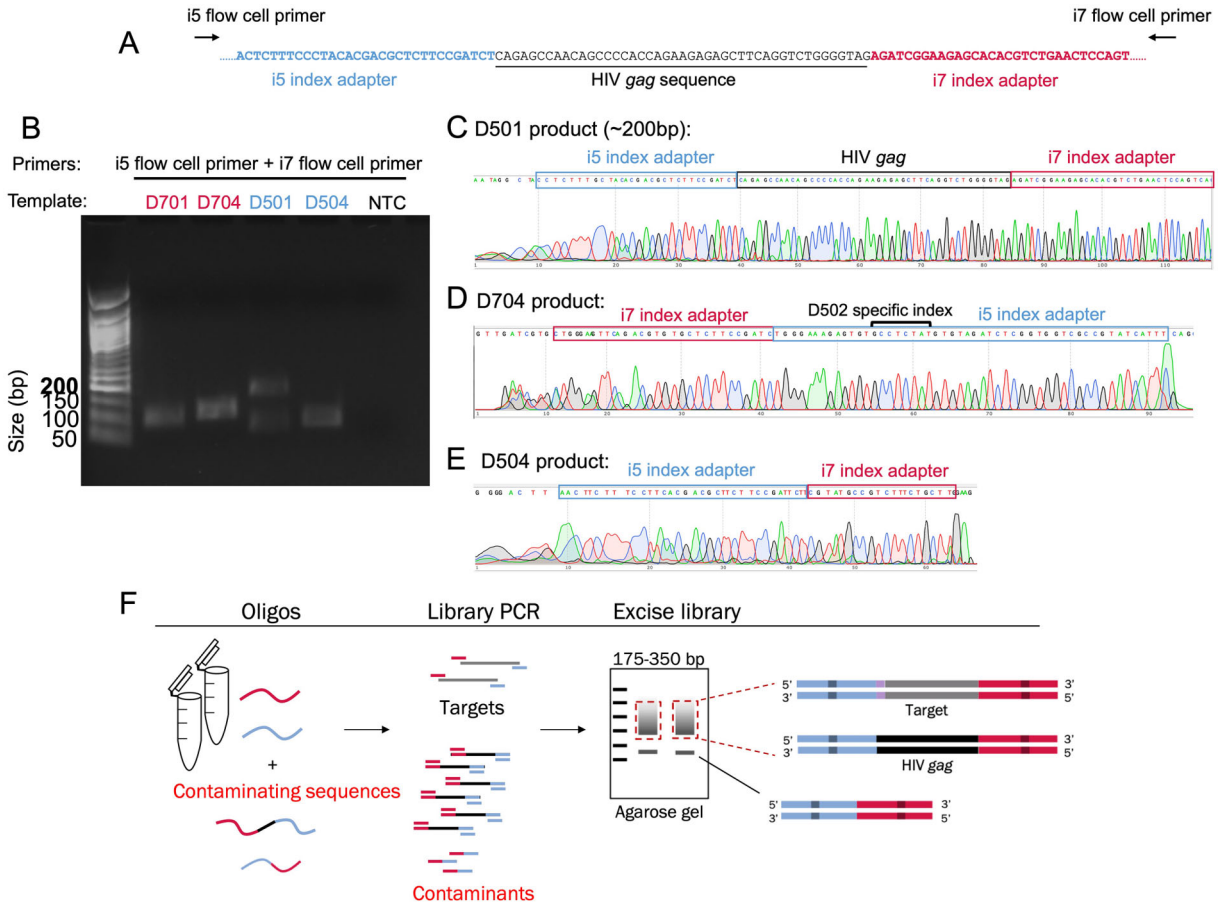


Figure 6. The contaminating HIV *gag* sequence is present between index adapter sequences as a functional library molecule in the supplied oligos. (**A**) Annotated seCLIP sequencing read of the HIV *gag* contaminant. The HIV *gag* sequence is shown in black and partial i5 and i7 index adapters in blue and red, respectively. Location of the primers used in PCR reactions shown in B are indicated as arrows. (**B**) Agarose gel electrophoresis of PCR reactions with primers against the flow cell binding sites of index adapters [see panel (A)] performed with i7 index adapters (red) or i5 index adapters (blue) as template or no template water control (NTC). (**C**) Sanger sequencing chromatogram of the ~200 bp gel-purified PCR product from B with i5 index adapter D501 used as template (lane 4). The 45nt contaminating HIV *gag* sequence and (partial) i5 and i7 index adapter sequences are annotated. (**D**) Chromatogram showing the Sanger sequencing result from the ~100 to 150 bp PCR product from B where i7 index adapter D704 was used as template (lane 3). Partial i5 and i7 index adapter sequences and the 8nt index specific to i5 index adapter D502 are indicated. (**E**) Same as C and D but from the PCR product with i5 index adapter D504 in B (lane 5). (**F**) Schematic image depicting how the library-structured contaminant, HIV *gag*, in commercial oligos was incorporated in the CLIP library that was sent for sequencing.

gag specific primers, a product that corresponds to the exact sequence and correct 'library structure' for NGS identified in our seCLIP data could be readily amplified from the commercially supplied oligo.

Moreover, analysis of the sequence of the other PCR products shown in Fig. 6B revealed that the HIV *gag* sequence was not the only unrelated sequence that was part of the index adapter. For example, from the reaction where i7 index adapter D704 was used as template (Fig. 6B and D) the sequence of i5 index adapter D502 could also be identified. Specifically, apart from the last 21 nucleotides at the 3′ end, the full sequence of D502 could be clearly identified from the sequencing data directly adjacent to the i7 index adapter sequence. Due to the specific 8 nt index corresponding to D502 this sequence could be assigned to this i5 index adapter compared to the other adapters. Thus, when we performed PCR with the two primers against the flow cell binding sequence of i5 and i7 index adapters, amplification was observed and subsequent sequencing revealed that the received i7 index adapter D704 was a mixture that also contains other irrelevant sequences within the same molecule (i.e. D502). A similar fusion of sequences within an oligo was also observed for i5 index adapter D504 (Fig. 6E). Here, the first 5′ nucleotides of an i7 index adapter (the annealing site for i7 flow cell primer) was flanked by the i5 index adapter sequence and thus we concluded that the sequence of this oligo corresponded not only to the ordered sequence but also partially to a (fragmented) sequence of an i7 index adapter.

The PCR products that were mixes of index adapters without the HIV *gag* contaminant were not identified in our seCLIP data as these ran below the cDNA library size cut-off (window 175–350 bp) (Fig. 6F). If they were present within the multiplexed library that was sent for sequencing, these reads would have been discarded during the adapter trimming stage of data processing as they only contain adapter sequences.

To summarize this case study, we identified an HIV *gag* sequence as a major contaminant in seCLIP libraries and through systematic investigation traced the source of this sequence back to the commercially supplied index adapter oligos that were used for the final library amplifications. Within the commercially supplied oligos, we found molecules with the

HIV *gag* between i5 and i7 index adapter sequences (Fig. 6F). Critically, this order of sequences formed a library structure that enabled it to be amplified and sequenced thus heavily contaminating the reads of the NGS and precluding a meaningful analysis of the correctly sequenced reads. In addition, other unrelated sequences were found in the index adapters in these oligos. We repeated the entire seCLIP experiment after replacing these oligos with newly ordered PAGE-purified oligos from a different commercial source, and no contaminating sequences were identified in the new data, confirming that these oligos were the unquestionable source of the contamination.

Perspectives

NGS-based techniques have now become the method of choice for a wide range of genomic and transcriptomic studies. As all these techniques require successive steps of introduction of synthetic oligos to RNA and DNA intermediates, and amplification of the resulting products, these protocols exquisitely depend on the purity and specificity of these oligos. Methods of oligonucleotide synthesis have vastly improved over the last decades and currently automated solid-phase synthesis has given rise to remarkable improvement in the quality and scale of oligonucleotide synthesis. Synthesized oligonucleotides are further purified by desalting, cartridge purification, PAGE, or HPLC, respectively, in order to achieve successively higher grades of purity and minimize sequence failures [15]. However, it is still possible to have commercially supplied oligos, especially index adapter oligos, from reputed suppliers which may be contaminated with unrelated sequences or failed synthesis products [16]. It is of importance for researchers using NGS-based approaches to be able to recognize the presence of such contaminating sequences in their data and detect the source of the contamination in the multiple sets of oligos which are used for NGS-based experiments.

We have therefore described here an unexpected source of experimental error that was caused by a profound cross-contamination of commercially supplied index adapter oligos. These contaminations resulted in the generation of uninterpretable data sets in an seCLIP experiment. This became apparent during data analysis of this highly complex and resource-intensive experiment. Systematic investigation into the cause of the experimental results resulted in the identification of contaminating index adapter oligos, which included all the required sequences to be carried over until the final step of NGS.

Although we exemplify the importance of identifying the source of the contamination in the interpretation of seCLIP data, the implications of the findings described here extend well beyond CLIP-seq and are directly relevant to a broad spectrum of high-throughput sequencing applications that critically depend on custom-synthesized oligonucleotides for library construction. Bulk RNA-seq, ChIP-seq, ATAC-seq, and DNA methylation profiling all rely on adapter and index oligos that are introduced at early or intermediate steps of library preparation, where even low-level contamination with amplification-competent, library-structured molecules can profoundly distort library complexity, duplication rates and quantitative signal. In targeted resequencing and amplicon-based workflows, such contaminants may preferentially amplify due to their favourable length and primer complementarity, leading to misleading enrichment and false-positive target detection.

Single-cell and single-nucleus sequencing approaches are particularly vulnerable to this class of artefact. Given their extreme sensitivity, limited input material and reliance on combinatorial indexing or barcoding strategies, contaminating oligos can propagate across large numbers of cells or droplets, manifesting as pervasive background signal, barcode collisions or spurious shared features between otherwise unrelated cells. In such settings, artefacts introduced at the oligo level may be incorrectly interpreted as biologically meaningful cell populations or technical batch effects.

Similarly, CRISPR-based pooled screening platforms, which employ guide RNA libraries and sequencing adapters in parallel, are susceptible to cross-contamination between unrelated oligo species. Here, the presence of unintended sequences embedded within functional adapter contexts could result in erroneous guide assignment or confounded genotype–phenotype associations. The systematic quality control framework presented in this study—integrating analysis of unmapped reads, structural analysis of sequencing artefacts, and direct interrogation of supplied oligonucleotides—provides a generally applicable strategy for identifying and tracing oligo-derived contaminants. Incorporating such validation steps into routine quality control workflows, particularly for adapter and index oligos used across multiple high-throughput platforms, may substantially improve the robustness and interpretability of NGS data across diverse experimental contexts. In addition, to increase the stringency of the production process and enhance the confidence in oligonucleotides used for research purposes, it may be advisable for researchers to engage directly with suppliers and share with them the data of their systematic troubleshooting approaches in case contaminating sequences are detected in supplied oligonucleotides. This will help suppliers to identify potential sources of cross-contamination or quality control failure in their manufacturing processes, and thereby adopt remedial steps to limit the chances of such contamination.

Specifically, it was found that the HIV *gag* contaminant sequence was flanked by full length NGS adapter oligo sequences but lacked a UMI. This ‘library-structure’ of the contaminant can explain the biased amplification of the contaminant at the expense of the cDNAs of interest, as normally these adapter oligos are only partially complementary to the intended targets. The PCR products resulting from the use of these adapter molecules were thus grossly over-represented at the step of data analysis.

In the particular case detailed here, the contamination was readily detected during data processing due to its non-human origin, which was highly dissimilar from the anticipated results. However, if this contaminant would have been of human origin, the identification of artefactual results might have been more challenging, possibly resulting in misinterpretation of experimental data such as incorrect RBP interaction target determinations. Crucially, the presence of the contaminating sequence(s) as part of the same molecule as the supplied adapter oligo sequences clearly demonstrate that the contaminating sequences were not introduced from the laboratory during any step of the experimental workflow. Interestingly, given the ‘library-structure’ and the size of the major contam-

inating oligo, it may have been designed by researchers, and synthesized at the same manufacturing facility, as it is reminiscent of a 'spike-in control'. Synthetic spike-in controls are RNA or DNA molecules that function as internal quality controls for NGS workflows and constitute sequences from different species to distinguish them from the sequences of interest [17]. Cross-contamination however does not explain how longer oligos composed of multiple adapter sequences were generated.

When tracking the source of the contamination, we were surprised to find that the quality issues of the commercially provided and supposedly rigorously quality-controlled oligos could be recognized by standard laboratory technology such as PAGE. While this report documents unexpected issues with the quality of oligos in the context of CLIP protocols, our experience does not appear to be a unique exception but a real world challenge with commercial vendors around the world [10]. As described here, problems with purity and integrity of oligos, despite assurances of rigorous quality control, can have a substantial impact on experimental results and cause waste of scientific resources and time. Furthermore, contamination of oligos used routinely for molecular diagnostic procedures, such as primers for RT-PCR, poses problems of possible misdiagnosis. Reports of contamination of SARS-CoV-2 detection primers during the COVID-19 pandemic by synthetic positive control oligonucleotides manufactured by the same supplier emphasize this point [18].

While we utilized PAGE-purified index adapter oligos in this case study in accordance with a widely used published protocol for seCLIP-seq [11], it is important to note that such protocols may not reflect current best practices in oligonucleotide manufacturing. In particular, PAGE purification, although historically employed to increase oligo purity, involves extensive post-synthesis handling steps that increase the risk of cross-contamination. This is in contrast to evolving manufacturer recommendations, which increasingly favour alternative purification strategies optimized for high-throughput sequencing applications. Indeed, both PAGE and, to a lesser extent, HPLC purification have been associated with contamination levels that can compromise sensitive downstream applications such as NGS and dPCR [16]. Modern manufacturer-recommended oligos, produced using improved chemistries and quality control processes, typically exhibit substantially lower cross-contamination rates (~0.01%–0.05%). We therefore recommend prioritizing manufacturer guidance over legacy protocols to minimize artefacts such as index misassignment, thereby improving data integrity and avoiding unnecessary experimental repetition.

While no standard protocol is available to detect similar contaminants in commercially supplied oligos, this case study provides a systematic methodological approach to detect contaminating sequences in commercially synthesized oligos designed for NGS library preparation. This required successive steps of excluding possible sources of the contamination, including from in-house and sequencing facility sources, and then investigating the multiple possible oligos which are utilized as inputs in the NGS workflow before being able to identify the particular oligo set which contained the contaminating oligos. We hope that this perspective will help the extensive community of researchers using NGS-based approaches to be able to recognize and detect the presence of possible contaminating sequences in their commercially sourced oligonucleotides.

Acknowledgements

We acknowledge EMBL's core facilities, specifically GeneCore, for their guidance and expert services. We also acknowledge EMBL IT Services for HPC resources and services.

Author contributions: Manouk Gerritsen (Formal analysis [equal], Investigation [equal], Methodology [equal], Visualization [equal], Writing—original draft [equal]), Partho S. Ray (Formal analysis [equal], Investigation [equal], Methodology [equal], Visualization [equal], Writing—original draft [equal]), Thileepan Sekaran (Formal analysis [lead]), Thomas Schwarzl (Formal analysis [supporting]), Matthias W. Hentze (Funding acquisition [equal], Supervision [equal]), and Andreas E. Kulozik (Funding acquisition [equal], Supervision [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

The authors declare no conflicts of interest.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) grant Ku563/17-1. M.G. was supported by a predoctoral fellowship from European Molecular Biology Laboratory Heidelberg. P.S.R. was supprted by Actions Thématiques Incitatives de Genopole (ATIGE) grant from Genopole, Evry, France.

Data availability

The raw NGS sequencing data have been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB87524.

References

1. Mohammed AA, AlShaer D, Al Musaimi O. Oligonucleotides: evolution and innovation. *Med Chem Res* 2024;33:2204–20. https://doi.org/10.1007/s00044-024-03352-7
2. Moccia M, Pascucci B, Saviano M *et al.* Advances in nucleic acid research: exploring the potential of oligonucleotides for therapeutic applications and biological studies. *Int J Mol Sci* 2023;25:146, https://doi.org/10.3390/ijms25010146 .
3. Kim H, Kim S, Lee D *et al.* Oligonucleotide therapeutics and their chemical modification strategies for clinical applications. *J. Pharm. Investig.* 2024;54:415–33. https://doi.org/10.1007/s40005-024-00669-8
4. Lee FCY, Ule J. Advances in CLIP technologies for studies of protein–RNA interactions. *Mol Cell* 2018;69:354–69. https://doi.org/10.1016/j.molcel.2018.01.005
5. Ule J, Jensen KB, Ruggiu M *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* 2003;302:1212–5. https://doi.org/10.1126/science.1090095
6. Hafner M, Katsantoni M, Köster T *et al.* CLIP and complementary methods. *Nat Rev Methods Primers* 2021;1:20. https://doi.org/10.1038/s43586-021-00018-1
7. Hauer C, Curk T, Anders S *et al.* Improved binding site assignment by high-resolution mapping of RNA–protein interactions using iCLIP. *Nat Commun* 2015;6:7921. https://doi.org/10.1038/ncomms8921

8. Sinha R, Stanley G, Gulati GS *et al.*. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. 2017. bioRxiv, https://www.biorxiv.org/content/10.1101/125724v1, 09 April 2017, preprint: not peer reviewed.
9. MacConaill LE, Burns RT, Nag A *et al.* Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics [Electronic Resource]* 2018;19:30. https://doi.org/10.1186/s12864-017-4428-5
10. Arakawa H, Miura H, Quadros RM *et al.* Cross-contamination of CRISPR guides and other unrelated nucleotide sequences among commercial oligonucleotides. *Nucleic Acids Res* 2024;52:3137–45. https://doi.org/10.1093/nar/gkae068
11. Blue SM, Yee BA, Pratt GA *et al.* Transcriptome-wide identification of RNA-binding protein binding sites using seCLIP-seq. *Nat Protoc* 2022;17:1223–65. https://doi.org/10.1038/s41596-022-00680-z
12. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 2015;6:11. https://doi.org/10.1186/s13100-015-0041-9
13. Bell NM, Lever AML. HIV Gag polyprotein: processing and early viral particle assembly. *Trends Microbiol* 2013;21:136–44. https://doi.org/10.1016/j.tim.2012.11.006
14. Van Nostrand EL, Nguyen TB, Gelboin-Burkhart C *et al.* Robust, cost-effective profiling of RNA binding protein targets with single-end enhanced crosslinking and immunoprecipitation (seCLIP). *Methods Mol Biol* 2017;1648:177–200.
15. Ellington A, Pollard JD, Jr. Synthesis and purification of oligonucleotides. *Curr Protoc Mol Biol* 2001;Chapter 2, Unit2.11. https://doi.org/10.1002/0471142727.mb0211s42
16. Max Leopold MH, Spanopoulou A, Dick R *et al.* A new oligonucleotide synthesis process for high quality oligonucleotides with superior performance in NGS applications. https://eurofinsgenomics.eu/media/1612446/applicationnote_ngs_oligos_20.pdf (5 December 2025, date last accessed).
17. Chen K, Hu Z, Xia Z *et al.* The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol Cell Biol* 2016;36:662–7. https://doi.org/10.1128/MCB.00970-14
18. Wang CYT, Buckley C, Bletchly C *et al.* Contamination of SARS-CoV-2 RT-PCR probes at the oligonucleotide manufacturer. *Pathology (Phila)* 2020;52:814–6. https://doi.org/10.1016/j.pathol.2020.08.002

Received: October 31, 2025. **Revised:** March 28, 2026. **Accepted:** March 31, 2026