Verification of CRISPR editing by Xdrop™ Indirect Sequence Capture followed by short- and long- read sequencing

Blondal Thorarinn[1], Gamba Cristina[1], Johnston Camille M.[1], Riising Eva M.[1], Mikkelsen Marie J.[1], Mouritzen Peter[1*]

[1] Samplix ApS, Mileparken 28, Herlev, Denmark

***Corresponding author:**
Name and Surname: Peter Mouritzen
E-mail address: pmo@samplix.com
Full postal address: Samplix ApS, Mileparken 28, Herlev, Denmark
Telephone: +45 40902107

Abstract

Validation of CRISPR-Cas9 editing typically explore the immediate vicinity of the target site and distal off-target sequences, which have led to the conclusion that CRISPR-Cas9 editing is very specific. However, an increasing number of studies suggest that on-target unintended editing events like deletions and insertions are relatively frequent but unfortunately often missed in the validation of CRISPR-Cas9 editing. The deletions may be several kb and only affect one allele. The golden standard in molecular validation of gene editing is direct sequencing of relatively short PCR amplicons. This approach will allow detection of smaller editing events but will fail to detect larger rearrangements in particular when these only affect one of the two alleles. Detection of larger rearrangements requires that an extended region is analyzed and the characterization of events may benefit from long-read sequencing. Here we present a new microfluidic technology Xdrop™ that allows targeted enrichment of longer regions (~ 100 kb) using just a standard PCR primer set. We sequenced the enriched region on long- and short- read sequencing platforms to unravel unknown and unintended genome edits resulting from CRISPR-Cas9 gene editing. We show how loss of heterozygosity, due to an accidental insertion in one of the alleles, remained undetected using standard procedures but was discovered implementing the Xdrop™ system. This demonstrates the potential of such technology in CRISPR gene editing assessments, providing an extremely accurate tool.

Keywords (max 6): Target Enrichment, Indirect Sequence Capture, CRISPR editing validation, long-read sequencing, loss of heterozygosity.

1. Introduction

The high precision, specificity and efficiency of CRISPR has provided an unprecedented improvement in genetic engineering with a performance spectacularly improved relative to former technologies like gene targeting [1,2]. The advancement has accelerated the extended exploitation of genetic engineering in organisms from plants and animals to humans. The considerations of applying CRISPR for gene therapy in humans have spurred concerns about safety where the most pressing questions currently relate to accuracy of the applied CRISPR editing and risk of potential off-target editing.

The concern about distal off-target events obviously relates to the potential consequences of undesired interruption of function for non-target genes elsewhere in the genome - in particular in genomic regions containing oncogenes and tumor suppressor genes [3]. Review of studies applying CRISPR-Cas9 in mouse and human cells suggests that distal off-target events are rare and undetectable [4]. The commonly applied strategy in studies investigating occurrence of distal off-targets events is to sequence bioinformatically identified top gRNA

off-target candidates. While this type of distal off-target events may be undetectable, other unintended editing events are occasionally detected including duplications, indels, single nucleotide variants (SNV), but also larger rearrangements [5]. Reintegration of larger excised regions in the vicinity of the edited region may occur when applying strategies using two gRNAs simultaneously [6]. However, strategies applying just a single gRNA may also produce unintended editing in the form of larger deletions and smaller indels [7]. Focused analysis directed towards a more extended region surrounding the site of gene editing suggests that on-target unintended editing may in fact not be infrequent [8]. The underlying mechanism responsible for this is the error-prone non-homologous end-joining (NHEJ) repair pathway induced after the Cas-9 nuclease generates double strand breaks (DSB). NHEJ dominates over the more precise homology-directed repair (HDR) pathway [9].

A multitude of methods are employed to verify correct gene editing and to ensure that on-target unintended editing events have not occurred. The most commonly applied molecular method is direct DNA sequencing of PCR amplicons, which are typically < 1000 bp in size [10–12]. Other common methodologies include Restriction Fragment Length Polymorphism (RFLP) analysis, cloning of PCR amplicons followed by Sanger sequencing, NGS deep sequencing of PCR amplicons, array-Comparative Genomic Hybridization (CGH), and FISH [6,7,9]. The work by Kosicki et al 2018 have underlined the importance of extending the molecular analysis to include several kilobases surrounding the site of CRISPR-Cas9 induced DSB and it demonstrates how long-read sequencing may ease the resolution of occurred rearrangements. Long-range PCR allowed Kosicki et al (2018) to generate long range sequence information but this is cumbersome to optimize and has an upper limit in the size of amplicon that can be generated.

Other approaches have been developed for targeted enrichment of long fragments: (1) Targeted locus amplification (TLA), based on physical crosslinks in adjacent sequences [13] , (2) CRISPR-Cas9 systems, where selected loci are identified using guide RNA [14,15], and (3) Region Specific Extraction (RSE), where a large selection of oligos is used for capturing target molecules [16]. All these techniques can take advantage of long-read sequencing [17–19] allowing for characterization of GC-rich regions and complex genomic regions, i.e. structural variants including insertion and deletions, as well as gene duplications, repetitive regions, haplotype phasing *and de novo* genome assembly [20–22]. However, the CRISPR-Cas9 system for targeted enrichment requires relatively high amount of DNA input in the order of micrograms and results in moderate enrichment and coverage of the target region [14,15] . TLA has a challenging implementation [23], and RSE requires design of several primers to cover larger regions and the enrichment is only around 100-fold [16].

Here we have applied a new microfluidic approach for enrichment (100 to > 1000-fold) of target regions of ~ 100 kb in size, based on Indirect Sequence Capture [24]. The enriched DNA is compatible with both short and long read sequencing. The technology, called Xdrop™, requires only the design of a single standard primer pair, which offers flexibility when targets frequently change e.g. in different gene editing experiments. Given the Indirect Sequence Capture approach, the primer set can be designed at kilobases away from the gene editing site. This provides a valuable advantage in detecting large deletions, which can be easily overlooked by conventional PCR screening strategies. Moreover, the two alleles can routinely harbor deletions of distinctly different sizes further complicating their identification [7]. Another advantage is that Xdrop™ enrichment only requires a limited amount of input DNA. On average only 10 ng of human DNA is needed.

Here we describe the application of the Xdrop™ technology in detecting unintended gene edits generated by CRISPR-Cas9 in a set of isogenic cell lines. Importantly, these unintended editing events had remained undetected by a comprehensive analysis using several standard approaches [25]. Apart from CRISPR-Cas9 verification sequencing to unravel unintended editing, Xdrop™ Indirect Sequence Capture technology's ability to enrich for long DNA fragments makes it highly applicable for other genomic analyses like gap closing, structural variations, tandem repeats, genome integrations.

1.1 Description of the Xdrop™ technology

The Xdrop™ technology for targeted enrichment is based on partitioning of the DNA into millions of picolitre reactions inside double emulsion droplets produced in the Xdrop™ instrument. Droplets containing the DNA of interest are identified by a small (~150 bp) amplicon (Detection Sequence) produced by droplet PCR. The droplets are stained with a fluorescent dye and only the droplets containing the DNA of interest will become fluorescent. Because of the size and composition of the PCR droplets, they can be sorted on a standard Fluorescence activated cell sorter enriching the droplets containing the DNA of interest. The enriched DNA is then isolated and amplified by multiple displacement amplification in droplets (dMDA). Thus, the Xdrop™ technology enables Indirect Sequence Capture and subsequent amplification of specific genomic *loci*. As mentioned above, this technology can be used for a wide variety of applications. In this study, we have employed the Xdrop™ technology to elucidate the editing events, intended as well as unintended, of CRISPR-Cas9 gene editing in human cell lines.

The Xdrop™ technology can enrich for long (~100 kb) target DNA regions while only requiring the design of a single primer pair to generate a short Detection Sequence (~150 bp) (Fig.1). This Detection Sequence is optimally placed in the centre of the kilobases-long Region of Interest (ROI) or in flanking known regions if the centre of the ROI is unknown. The Detection Sequence amplicon is exclusively used for detection, selection and enrichment of the full-length ROI, which is ultimately captured and sequenced. Being able to select for long fragments based on a short Detection Sequence means that both known and unknown genomic regions can be indirectly targeted, the latter by designing Detection Primers in flanking regions.

More specifically, Xdrop™ Indirect Sequence Capture combines (1) high-resolution droplet PCR (dPCR) with (2) droplet sorting and (3) Multiple Displacement Amplification in droplets (dMDA). Steps (1) and (3) involve the use of the Xdrop™ instrument and a standard thermal cycler, while for step (2) a standard cell sorter is needed (Fig.1).

The Xdrop™ workflow (Fig. 1) starts with high molecular weight (HMW) genomic DNA and ends with target enriched long DNA fragments ready for sequencing. To perform this workflow, the Xdrop™, dPCR cartridges, dMDA cartridges and associated reagents are needed (Samplix, Denmark). For a full list of items see Appendix A. Other relevant equipment, consumables and reagents are listed in Appendix B.

2. Material and methods

2.1 Samples

The iPSC line BIONi010-C and four CRISPR-Cas9 gene edited APOE derivatives were described in [25]. Three of these cell lines were edited at single nucleotide positions in Exon 4 to modify genotypes at two different SNPs (rs429358 and rs7412), recognized as the most common genetic risk factor for sporadic Alzheimer's disease (AD). The editing was performed to obtain isogenic cell lines homozygous at the two SNP positions namely APOE-ε4/ε4 (rs429358 C/C, rs7412 C/C), APOE-ε3/ε3 (rs429358 T/T, rs7412 C/C) and APOE-ε2/ε2 (rs429358 T/T, rs7412 T/T). The APOE-ε2 (rs429358 T, rs7412 T) and APOE-ε4 (rs429358 C, rs7412 C) haplotypes are associated with lower and higher risk of AD, respectively [26,27]. The cell lines were developed to serve as an AD model system to be able to study

the biological mechanisms by which the APOE alleles affect the risk of developing AD [25]. Moreover, an APOE knock-out cell line was generated, carrying an 8-nucleotide deletion in Exon 2. All cell lines were generated by editing the iPSC line BIONi010-C (genotype APOE-ε3/ε4), previously generated from human skin fibroblasts [28], as described apart from APOE-ε2/ε2 which was generated from the APOE-ε3/ε3 cell line [25].

We analysed the APOE locus and surrounding region in all five cell lines but included in this study only data relative to those two cell lines where we detected unintended edits (APOE-ε2/ε2 and APOE-ε3/ε3).

2.2 Target enrichment by Xdrop™

2.2.1 Input DNA

High molecular weight (HMW) DNA was isolated from thawed cell lines by Bioneer A/S using the Blood & Cell Culture DNA Mini Kit (Qiagen). Around 300 ng of the isolated genomic DNA were further subjected to capture bead clean-up step utilizing HighPrep™ PCR Clean-up Capture bead system (MagBio Genomics Inc.) in 1:1 (Vol:Vol) ratio according to the manufacturer's instructions. HMW DNA was resuspended in 10 mM Tris pH 8 and quantified by Quantus™ Fluorometer (Promega Inc). The DNA size and integrity were assessed on TapeStation™ instrument (Agilent, Genomic DNA ScreenTape), and 10 ng of HMW DNA was used as input to the Xdrop™ workflow.
In order to obtain a significant enrichment, we calculated the optimal amount of input template DNA using the automated online tool "DNA input calculator" available online at samplix.com.

2.2.2 Detection Sequence design and testing

Using the "Primer design tool" available online at samplix.com, we designed the Detection primer set for dPCR (fw: GCTAGCCGTCGATTGGAGAA, rev: CATCTCAGTCCCAGTCTCGC, amplicon position: chr19: 44,906,188 – 44,906,317) ~ 2kb upstream the CRISPR-Cas9 edited region. We also designed a second primer set, the Validation Primer set (fw: GGAGGCCTCCGTTTTCTCAA, rev: CCGGACTTAAGGCAGCATCA, amplicon position: chr19: 44,903,692 – 44,903,842) to validate the enrichment of the ROI at the end of the Xdrop™ workflow by qPCR. Both Detection and Validation Primers were tested prior to the enrichment by qPCR, using as input a representative HMW DNA sample and reagents from Samplix Primer test PCR kit (Cat. No. RE10200), following manufacturers' recommendations.

2.2.3 Target detection by dPCR

Double emulsion droplets were generated in the Xdrop™ instrument by using the dPCR cartridge, where all samples were run in parallel. The dPCR cartridge was loaded with the 1x dPCR buffer, dPCR mix, and dPCR oil (in that order, to ensure successful droplet production) and sealed to avoid cross-contamination, following manufacturers' recommendations.

After droplet generation, droplets were collected, aliquoted in four 0.2 ml PCR tubes and transferred to a thermal cycler, where the following program with slow ramping at all stages

(1,5°C/sec) was run: 30°C (5 sec), 94°C (3 min), 40 cycles of 94°C (3 sec) and 60°C (30 sec), 4°C (∞). Droplets were stored overnight at 4°C before flow cytometer sorting.

2.2.4 Flow cytometry sorting of target

A SONY SH800S cell sorter with 100 µm nozzle sorting chip was used to sort the positive dPCR droplets (containing the ROI) from the negative droplet population. The dPCR droplets were stained with Droplet dye before sorting following manufacturers' instructions. The Samplix Cell sorter control kit, consisting of ready-made dPCR droplets with a large population of positive droplets (10-15% of total dPCR droplets), was used to easily pre-set flow cytometry sorting gates, and sample pressure was set to aim for ~ 5000 events/sec. Positive fluorescent droplets represented ~ 0.01% of total dPCR droplets and could be clearly separated form negative droplets, gated and sorted out into 15 µL of PCR grade water in a 1.5 ml DNA LoBind tube (Eppendorf).

The expected number of positive droplets correlated with the number calculated using the "Enrichment Predictor" tool available online at samplix.com. After sorting, the positively sorted dPCR droplets were kept for a short time at 4°C before dMDA amplification. The dMDA has to be performed on the same day as the droplet sorting.

2.2.5 Droplet Multiple displacement amplification (dMDA)

Immediately after sorting, the DNA was released from the positive droplets by adding Break Solution and Break colour to each sample. All of the clear bottom phase (Break solution) was collected and discarded. The dMDA reactions were then set up using the reagents from Samplix dMDA kit and following manufacturers' recommendations. To monitor contamination, we included as negative controls both the sheath fluid from the flow cytometer and PCR grade water. As a positive control we used 1 pg Human Genomic DNA, Female (Part n. G152A, Promega Corp.). The samples and the dMDA reagents were mixed and injected into the dMDA cartridge followed by overlaying with oil. The loading of reagents in the dMDA cartridge carefully followed manufacturers' recommendations. The sealed cartridge was inserted into the Xdrop™ instrument for single emulsion droplets generation. Then droplets were collected and transferred into a 0.2 ml PCR tube and incubated in a thermal cycler for 16 hours at 30°C, where the reaction was terminated at 65°C (10 min) and then kept at 4°C (∞).
After dMDA incubation, DNA was harvested by adding Break Solution and Break colour to each sample and then the clear bottom phase (Break solution) was collected and discarded.

2.2.6 Evaluation of enrichment

The total amount of enriched DNA released from the dMDA droplets was measured with Quantus™ Fluorometer (Promega Inc.) and fragment size distribution was inspected with a TapeStation™ (Agilent, Genomic DNA ScreenTape), following manufacturers' instructions. Fold enrichment of target DNA was firstly assessed by qPCR using the Validation Primers, not overlapping with the Detection Sequence (amplified by the Detection Primers). Assessment by qPCR, which provides a rough estimate of the enrichment success prior to sequencing, was performed by following manufacturers' instructions. Calculation of fold enrichment were assessed with the online "Enrichment calculator" available at samplix.com.

2.3 Library preparation and sequencing

2.3.1 Short-read sequencing

For short-read sequencing library generation, 300 ng of enriched DNA samples from the Xdrop™ workflow were fragmented using the NEBNext® Ultra™ II FS DNA Library Prep Kit (New England Biolabs) according to the manufacturer's recommendation with 5 minutes incubation time. After fragmentation, the samples were processed according to the TruSeq PCR Free Library Preparation protocol starting with end-repair, size selection, followed by dA tailing and barcoding and adaptor ligation as described by the manufacturer (Illumina Inc.). The size distribution of libraries was evaluated by Tapestation™ System (Agilent, D1000 ScreenTape) according to the manufacturer's instructions. Concentrations were calculated and the samples pooled in equimolar concentrations. Finally, the sample pool was quantified using KAPA Library Quant Kit – Illumina for BioRad (KAPA Biosystems) according to manufacturer's instructions. Short-read sequencing was performed on iSEQ100 desktop sequencer according to the manufacturer's instructions (Illumina Inc.) using paired end 2 x 151 protocol. Basecalling was performed automatically using GenerateFASTQ - 2.0.0 module.

2.3.2 Long-read sequencing

For long-read sequencing library generation, 1.5 µg of enriched DNA samples from the Xdrop™ workflow were used to generate Oxford Nanopore long-read sequencing libraries. First the amplified DNA was subjected to debranching, by incubation with15 units of T7 endonuclease I for 15 minutes at 37°C (New England biolabs) to reduce the branching introduced during the dMDA reaction. The DNA samples were DNA repaired and end-prepped using the NEBNext FFPE DNA Repair Mix (New England biolabs) and NEBNext End repair/dA-tailing Module (New England biolabs), and then barcoded using the Native Barcoding kit (Oxford Nanopore) followed by adaptor ligation by the Ligation Sequencing kit (Oxford Nanopore). All procedures were carried out following manufacturers' instructions and Samplix recommendations. Freshly prepared libraries were quantified by Quantus™ Fluorometer (Promega Corp.) according the manufacturer's recommendations and 30-50 fmol DNA loaded into a MinION flow cell (R9.4.1, Oxford Nanopore) and run on standard setting for 16 hours. Base calling of native reads was performed using Guppy v. 3.4.5 with high accuracy and quality filtering (Oxford Nanopore).

2.4 Data analysis

2.4.1 Long-reads data

Mapping against the human genome GRCh38.1, as well as the reconstructed hybrid sequence of the human genome (including the unintended insertion from the modified pEasyFlox plasmid), was performed using Minimap 2.0 [29] with default settings. Coverage was calculated using a combination of SAMtools [30] and BEDTools [31], and visualized in R [32]. Mapped reads were visualised using IGV 2.8.2 [33].

2.4.2 Short-reads data

Illumina reads were trimmed with Trimmomatic [34] using default settings and mapped to the construct and the human genome GRCh38.1 with BWA-MEM [30]. Coverage was calculated using SAMtools and BEDTools [31], and visualized in R [32].

2.4.3 Enrichment calculation

Sequencing-based enrichment was calculated for both short- and long-reads data using the following equation: Enrichment = (no. target reads / total mapped reads to reference) / (Target size / reference genome size) = (% reads mapped to target region) / (% target region of reference).

3. Results

3.1 Enrichment of edited APOE gene region

We enriched the APOE gene region from all five cell lines (see **Fig. 2**) using Xdrop™ (explained in detail in section 1.1), but report here only the results relative to cell lines APOE-ε2/ε2 and APOE-ε3/ε3, containing unintended edits. The Xdrop™ workflow provided a total of ~1.5 µg of enriched DNA, starting from 10 ng of HMW DNA. Evaluation of enrichment by qPCR estimated an enrichment of 1000-1800 fold. After enrichment, we proceeded with both long- and short-read sequencing, on Oxford Nanopore and Illumina platforms, respectively. Around 150,000 long reads (Oxford Nanopore) were obtained per sample and enrichment estimations on this data was ~1000-fold for the APOE gene and the 10 kb surrounding the Detection Sequence, and ~200-fold for the 100 kb around it (see Table 1). Around 5 million paired-end reads of 151 bp (Illumina) were obtained per sample and enrichment estimations on this data was ~700-fold for the APOE gene and the 10 kb surrounding the Detection Sequence, and ~200-fold for the 100 kb around it (see Table 1), in agreement with Oxford Nanopore data.

3.2 Detection of unintended edits in APOE Exon 4

All four cell lines edited with CRISPR-Cas9 technology were originally tested thoroughly by standard characterization approaches for validation of CRISPR gene editing, providing evidence of successful editing [25]. Three of these cell lines were edited at single nucleotide positions in Exon 4 to modify genotypes at two different SNPs (rs429358 and rs7412), associated with risk factors for sporadic AD and a fourth edited with an 8 bp deletion in Exon 2. Schmid et al. [25] used five complementary technologies to assess for correct gene editing, including genotyping and locus integrity assessment. Genotyping relied on Sanger sequencing of a short amplicon generated across the two SNPs in Exon 4 (rs429358 and rs7412 are only 139 bp apart) as well as another short amplicon generated across the deletion in Exon 2 in the KO cell line [25].

We could verify integrity and expected genotypes using the Xdrop™ enrichment workflow for 3 of the 5 cell lines (the parental APOE-ε3/ε4, APOE-ε4/ε4 and APOE-KO). However, in two of the five cell lines (APOE-ε2/ε2 and APOE-ε3/ε3) we could detect an unintended insertion of ~3.4 kb around the edited region of the haplotype APOE-ε4 (rs429358 C; rs7412 C). The unintended insertion caused a 5 bp deletion at positions 44,908,823 - 44,908,827 in chromosome 19 that was replaced by a 3,427 bp fragment from a custom pEasy Flox plasmid [35] modified with the insertion of a fragment of the human MAPT gene (Dr. Bjørn Holst (Bioneer), personal communication). The plasmid provided the cells with a resistance for neomycin to enable selection of nucleofected cells and was not supposed to be present in the final cell lines, nor to be integrated, which was tested by PCR [25].

The cell line APOE-ε3/ε3 (initially genotyped as rs429358 T/T; rs7412 C/C) was constructed from the original cell line BIONi010-C APOE-ε3/ε4 (rs429358 T/C; rs7412 C/C), while cell line APOE-ε2/ε2 (initially genotyped as rs429358 T/T; rs7412 T/T) was constructed from APOE-ε3/ε3 (Fig. 2). The reason why the identical insertion was detected in both alleged APOE-

ε3/ε3 and APOE-ε2/ε2 cell lines is that, most probably, this insertion happened in the haplotype APOE-ε4 (rs429358 C; rs7412 C) when editing APOE-ε3/ε3 from APOE-ε3/ε4, and was afterwards transmitted to the derived alleged APOE-ε2/ε2 (Fig. 2).

Interestingly, this unintended insertion remained undetected in the original study (Schmid et al. 2019) reporting on the generation of the set of cell lines and verifying the performed editing by several different methods. This probably occurred because the desired editing was to create two homozygous SNPs so failure in the ability to analyse one of the two alleles would suggest that both SNPs were indeed homozygous. The insertion did in fact only affect one of the two haplotypes (APOE-ε4: rs429358 C; rs7412 C), impairing the detection of it but leaving the other haplotype (APOE-ε3: rs429358 T; rs7412 C) intact to be detected. In the undetected haplotype, the primers for the amplicon encompassing the two SNPs went from a distance of 227 bp to a distance of > 3.5 kb, hampering the amplification of the affected haplotype due to the extended length, which is incompatible with standard PCR amplification (Fig. 3). In consequence, the alleged edited APOE-ε3/ε3 cell line appeared to have the correct rs429358 T/T and rs7412 C/C genotype, derived only from the intact haplotype (APOE-ε3: rs429358 T; rs7412 C). For the alleged APOE-ε2/ε2 cell line, editing did proceed correctly in the allele not affected by the insert making this appear to have the correct rs429358 T/T and rs7412 T/T genotype.

4. Discussion

In this study we have demonstrated that standard PCR based procedures for assessing CRISPR-Cas9 genome editing events do not provide a sufficiently accurate tool. A more truthful assessment is needed to detect unintentional editing, such as the one provided by the Xdrop™ workflow, which allows a thorough investigation of a ~100 kb long region surrounding the target site. The ability to perform both long- and short-read sequencing was an advantage in this study and should be in all instances where larger rearrangements have occurred. It has previously been suggested that CRISPR-Cas9 editing procedures can induce complex rearrangements, deletions and chromosomal truncations around the target region [8,9,36]. This can have significant consequences, especially affecting health and pharmaceutical related fields, such as pre-clinical model generation and gene therapy but also in basic sciences where undetected rearrangements may lead to drawing the wrong conclusions from experiments. In Kosicki et al. [8], long-range PCR amplicons were developed followed by long read nanopore sequencing to show deletion several kb in size in the immediate vicinity of the site of gene editing. However, the development and optimization of long-range-PCR can be time consuming and occasionally impossible. The Xdrop™ enrichment workflow only requires a single standard PCR primer set (Detection Primer set). This primer set can be flexibly designed in any region 5-10 kb away from the gene editing site which simplifies CRISPR verification sequencing.

The Xdrop™ enrichment of genomic regions of interest combined with both long and short read sequencing provided a broad overview of the region edited, allowing to investigate ~100 kb surrounding the edited region with an average enrichment of ~ 200 x (see Table 1 and Fig. 4). This allowed to identify and disentangle an unintended insertion that was completely missed with standard verification procedures [25].

This study highlights the great potential of the Xdrop™ technology in CRISPR-Cas9 gene editing assessments, including detection of unintended modifications that other technologies are currently missing with potentially critical consequences for subsequent outcome of gene therapy or for the biological conclusions made in model studies. The detected unintended edits reported here are on-target edits, however Xdrop™ may also be applied for enrichment of bioinformatically identified candidate distal off-targets such as tumour suppressor genes or

for unbiased analysis of insert integration patterns whether mediated by CRISPR, virus or other genetic engineering technologies.

Table 1. Sequencing statistics of the* Xdrop™ *enrichment of the APOE gene region*. RoT10 and RoT100 refer to the 10 kb and 100 kb region, respectively, surrounding the Detection amplicon* (i.e. droplet PCR amplicon, used for selection the positive droplets). ONT = Oxford Nanopore Technology; ILL = Illumina.

Statistic	E3/E3 (ONT)	E2/E2 (ONT)	E3/E3 (ILL)	E2/E2 (ILL)
APOE gene mean coverage (fold)	37	50	93	117
APOE enrichment (fold)	1,185	1,196	537	795
RoT10 enrichment (fold)	814	1,114	506	795
RoT100 enrichment (fold)	197	213	154	276
Total reads	138,156	168,079	5,222,516	4,433,588
Total reads average length (bp)	4,491	5,043	151	151
Total mapped reads	132,348	162,586	5,060,426	4,156,559
Percentage mapped reads to reference (%)	95.08	96.73	96.90	93.75
APOE gene mapped reads	175	217	3034	3689
RoT10 mapped reads	333	560	8,097	10,216
RoT100 mapped reads	804	1,070	24,290	35,423

Tips and Tricks

- Recommended input material for the Xdrop™ workflow is HMW DNA (> 50 kb) of high purity (260/280 and 260/230 ratios of ~1.8 and ~2.0, respectively)
- Bead purification of the HMW DNA sample can reduce noise and improve identification and sorting of positive droplets
- Detection Regions can be combined in multiplex reactions to extend the breadth of coverage of the Region of Interest
- Several online tools for primer design, input calculation and enrichment assessment are available at samplix.com
- Simultaneous design and test of several primer sets allows to choose for the best performing based on qPCR results
- The order and modality of loading reagents in Xdrop™ cartridges, extensively described in Samplix manual, is highly relevant to ensure a successful workflow
- Low binding plastic tubes help avoiding DNA loss, which is particular important for those steps where material is limited, such as when handling sorted positive droplets
- It is important to avoid contamination (i.e. working in a laminar flow hood, using filtered pipette tips, etc.) when setting up the dMDA reaction
- After sorting the positive droplets, it is crucial to keep samples on ice and proceed to the dMDA setup within 8 hours.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 848497. We would like to thank Dr. Bjørn Holst and Dr. Benjamin Schmid from Bioneer A/S for their help in data interpretation.

Data Accessibility

The reconstructed APOE reference (Chr19:44,900,000-44,920,000 including pEasy Flox fragment) has been submitted as Mendeley Data. Sequencing reads (Illumina and Oxford Nanopore) for cell lines *APOE- ε2/ε2 and APOE-ε3/ε3* mapping to the reconstructed reference have been submitted to the Genome Sequence Archive (GSA) database under the BioProject PRJCA002649.

Figures Legends

***Figure 1. Indirect Sequence Capture and unbiased amplification with Samplix* Xdrop™**.*The* Xdrop™ *workflow includes Indirect Sequence Capture in double emulsion droplets (dPCR box, blue) and multiple displacement amplification in single emulsion droplets (dMDA box, green). In the first step (1), the sample is mixed with the Detection Primers and PCR reagents and this mix (2) is partitioned into millions of double emulsion droplets of ~20 µm diameter using the Xdrop™ instrument and the dPCR cartridge (~40 minutes). These droplets are highly stable and are suitable for standard PCR cycling, flow cytometry analysis and sorting. (3) Droplets containing the Region of Interest are identified by the Detection sequence, a short amplicon (~150 bp) placed within or adjacent to the Region of Interest. Droplets are stained with an intercalating dye (4) and positive droplets are (5) identified by their fluorescence and physically separated from negative droplets using a standard cell sorter. DNA is released from double emulsion droplets (6) resulting in a population of long single DNA molecules enriched for the ROI and comprising kilobases of information. For downstream DNA amplification, (7) each long fragment derived from the enrichment is partitioned as single molecules into thousands of single emulsion ~85 µm diameter droplets (in less than one minute) for (8) high fidelity multiple displacement amplification in droplets (dMDA). Amplified enriched fragments around the region of interest are then (9) released from the single emulsion droplets. The Xdrop™ enriched DNA is compatible with (10) downstream analyses, such as long- and short-read sequencing.*
Figure 2. Gene editing workflow of the enriched cell line (from [25])***.*** *Five (supposedly) isogenic cell lines were created from BIONi010-C (genotype APOE-ε3/ε4), including the editing of the two pathogenic SNPs (rs429358 and rs7412) and a knockout (KO). The star (*) highlight the edited SNPs.*

Figure 3. Xdrop™ identified unintended insertion in the APOE gene of the* APOE-ε3/ε3 cell line. The zoom-in below Exon 4 shows the alleles detected after* Xdrop™ *enrichment: the Expected Allele (APOE-ε3/ε3 rs429358 T; rs7412 C) and the Unexpected Allele (rs429358 C; rs7412 C), including the 3.4 kb insertion of a portion of a modified pEasy Flox vector between positions Chr19: 44,908,822 - 44,908,828 (deleting 5 bases in the process). The red arrow above the APOE gene between Exon 1 and 2 shows the location of the Detection Sequence used for the dPCR enrichment assay.*

Figure 4. Depth of coverage of Illumina (blue) and Oxford Nanopore (green) sequencing data (sample ε2/ε2). *The main graph focuses on the 100 kb surrounding the Detection Sequence (low bar, red) designed to capture the APOE gene region (low bar, light blue). On the top right corner, a zoomed in area on the central 10 kb surrounding the Detection Sequence (RoT10).*

Figure 5. Alignment of reads to the APOE gene region. *IGV 2.8.2* [33] *screenshot of Illumina (ILL) and Oxford Nanopore Technology (ONT) alignment data to the Exon 4 of the APOE gene (including SNPs rs429358 [Gray = T ; Blue = C] and rs7412 [Gray = C; Red = T] are shown for cell lines APOE- ε2/ε2 and APOE-ε3/ε3.The insertion breakpoint is highlighted in red, where the drop of coverage is visible, due to the insertion of the plasmid in the correspondence of the APOE-ε4 haplotype (rs429358C; rs7412C).*

References

[1] M. Adli, The CRISPR tool kit for genome editing and beyond, Nat. Commun. (2018). https://doi.org/10.1038/s41467-018-04252-2.

[2] M.R. Capecchi, Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century, Nat. Rev. Genet. 6 (2005) 507–512. https://doi.org/10.1038/nrg1619.

[3] S.Q. Tsai, Z. Zheng, N.T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A.J. Iafrate, L.P. Le, M.J. Aryee, J.K. Joung, GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases, Nat. Biotechnol. (2015). https://doi.org/10.1038/nbt.3117.

[4] L. You, R. Tong, M. Li, Y. Liu, J. Xue, Y. Lu, Advancements and Obstacles of CRISPR-Cas9 Technology in Translational Research, Mol. Ther. - Methods Clin. Dev. (2019). https://doi.org/10.1016/j.omtm.2019.02.008.

[5] X. Niu, W. He, B. Song, Z. Ou, D. Fan, Y. Chen, Y. Fan, X. Sun, Combining single strand oligodeoxynucleotides and CRISPR/Cas9 to correct gene mutations in β-thalassemia-induced pluripotent stem sells, J. Biol. Chem. (2016). https://doi.org/10.1074/jbc.M116.719237.

[6] K. Boroviak, B. Fu, F. Yang, B. Doe, A. Bradley, Revealing hidden complexities of genomic rearrangements generated with Cas9, Sci. Rep. (2017). https://doi.org/10.1038/s41598-017-12740-6.

[7] H.Y. Shin, C. Wang, H.K. Lee, K.H. Yoo, X. Zeng, T. Kuhns, C.M. Yang, T. Mohr, C. Liu, L. Hennighausen, CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome, Nat. Commun. (2017). https://doi.org/10.1038/ncomms15464.

[8] M. Kosicki, K. Tomberg, A. Bradley, Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements, Nat. Biotechnol. 36 (2018) 765–771. https://doi.org/10.1038/nbt.4192.

[9] G. Cullot, J. Boutin, J. Toutain, F. Prat, P. Pennamen, C. Rooryck, M. Teichmann, E. Rousseau, I. Lamrissi-Garcia, V. Guyonnet-Duperat, A. Bibeyran, M. Lalanne, V. Prouzet-Mauléon, B. Turcq, C. Ged, J.-M. Blouin, E. Richard, S. Dabernat, F. Moreau-Gaudry, A. Bedel, CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations, Nat. Commun. 10 (2019) 1136. https://doi.org/10.1038/s41467-019-09006-2.

[10] H. Koike-Yusa, Y. Li, E.-P. Tan, M.D.C. Velasco-Herrera, K. Yusa, Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library, Nat. Biotechnol. 32 (2014) 267–273. https://doi.org/10.1038/nbt.2800.

[11] M. van Overbeek, D. Capurso, M.M. Carter, M.S. Thompson, E. Frias, C. Russ, J.S. Reece-Hoyes, C. Nye, S. Gradia, B. Vidal, J. Zheng, G.R. Hoffman, C.K. Fuller, A.P. May, DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks, Mol. Cell. 63 (2016) 633–646. https://doi.org/10.1016/j.molcel.2016.06.037.

[12] E.-P. Tan, Y. Li, M. Del Castillo Velasco-Herrera, K. Yusa, A. Bradley, Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells, Genesis. 53 (2015) 225–236. https://doi.org/10.1002/dvg.22835.
[13] P.J.P. de Vree, E. de Wit, M. Yilmaz, M. van de Heijning, P. Klous, M.J.A.M. Verstegen, Y. Wan, H. Teunissen, P.H.L. Krijger, G. Geeven, P.P. Eijk, D. Sie, B. Ylstra, L.O.M. Hulsman, M.F. van Dooren, L.J.C.M. van Zutven, A. van den Ouweland, S. Verbeek, K.W. van Dijk, M. Cornelissen, A.T. Das, B. Berkhout, B. Sikkema-Raddatz, E. van den Berg, P. van der Vlies, D. Weening, J.T. den Dunnen, M. Matusiak, M. Lamkanfi, M.J.L. Ligtenberg, P. ter Brugge, J. Jonkers, J.A. Foekens, J.W. Martens, R. van der Luijt, H.K.P. van Amstel, M. van Min, E. Splinter, W. de Laat, Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping, Nat. Biotechnol. 32 (2014) 1019–1025. https://doi.org/10.1038/nbt.2959.
[14] M.T.W. Ebbert, S.L. Farrugia, J.P. Sens, K. Jansen-West, T.F. Gendron, M. Prudencio, I.J. McLaughlin, B. Bowman, M. Seetin, M. Dejesus-Hernandez, J. Jackson, P.H. Brown, D.W. Dickson, M. Van Blitterswijk, R. Rademakers, L. Petrucelli, J.D. Fryer, Long-read sequencing across the C9orf72 “GGGGCC” repeat expansion: Implications for clinical use and genetic discovery efforts in human disease, Mol. Neurodegener. (2018). https://doi.org/10.1186/s13024-018-0274-4.
[15] T. Gabrieli, H. Sharim, D. Fridman, N. Arbib, Y. Michaeli, Y. Ebenstein, Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH), Nucleic Acids Res. (2018). https://doi.org/10.1093/nar/gky411.
[16] J. Dapprich, D. Ferriola, K. Mackiewicz, P.M. Clark, E. Rappaport, M. D’Arcy, A. Sasson, X. Gai, J. Schug, K.H. Kaestner, D. Monos, The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity, BMC Genomics. 17 (2016) 486. https://doi.org/10.1186/s12864-016-2836-6.
[17] A.S. Mikheyev, M.M.Y. Tin, A first look at the Oxford Nanopore MinION sequencer, Mol. Ecol. Resour. 14 (2014) 1097–1102. https://doi.org/10.1111/1755-0998.12324.
[18] M.O. Pollard, D. Gurdasani, A.J. Mentzer, T. Porter, M.S. Sandhu, Long reads: their purpose and place, Hum. Mol. Genet. 27 (2018) R234–R241. https://doi.org/10.1093/hmg/ddy177.
[19] A. Rhoads, K.F. Au, PacBio Sequencing and Its Applications, Genomics. Proteomics Bioinformatics. 13 (2015) 278–289. https://doi.org/10.1016/j.gpb.2015.08.002.
[20] K. Nakano, A. Shiroma, M. Shimoji, H. Tamotsu, N. Ashimine, S. Ohki, M. Shinzato, M. Minami, T. Nakanishi, K. Teruya, K. Satou, T. Hirano, Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area, Hum. Cell. 30 (2017) 149–161. https://doi.org/10.1007/s13577-017-0168-8.
[21] A. Ritz, A. Bashir, S. Sindi, D. Hsu, I. Hajirasouliha, B.J. Raphael, Characterization of structural variants with single molecule and hybrid sequencing approaches, Bioinformatics. 30 (2014) 3458–3466. https://doi.org/10.1093/bioinformatics/btu714.
[22] M. Wang, C.R. Beck, A.C. English, Q. Meng, C. Buhay, Y. Han, H. V. Doddapaneni, F. Yu, E. Boerwinkle, J.R. Lupski, D.M. Muzny, R.A. Gibbs, PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations, BMC Genomics. 16 (2015) 214. https://doi.org/10.1186/s12864-015-1370-2.
[23] A. Slesarev, L. Viswanathan, Y. Tang, T. Borgschulte, K. Achtien, D. Razafsky, D. Onions, A. Chang, C. Cote, CRISPR/Cas9 targeted CAPTURE of mammalian genomic regions for characterization by NGS, Sci. Rep. 9 (2019) 3587. https://doi.org/10.1038/s41598-019-39667-4.
[24] E.B. Madsen, T. Kvist, I. Höijer, A. Ameur, M.J. Mikkelsen, Xdrop: targeted sequencing of long DNA molecules from low input samples using droplet sorting, BioRxiv. (2018). https://doi.org/10.1101/409086.
[25] B. Schmid, K.R. Prehn, N. Nimsanor, B.I.A. Garcia, U. Poulsen, I. Jørring, M.A. Rasmussen, C. Clausen, U.A. Mau-Holzmann, S. Ramakrishna, R. Muddashetty, R.

Steeg, K. Bruce, P. Mackintosh, A. Ebneth, B. Holst, A. Cabrera-Socorro, Generation of a set of isogenic, gene-edited iPSC lines homozygous for all main APOE variants and an APOE knock-out line, Stem Cell Res. 34 (2019) 101349. https://doi.org/10.1016/j.scr.2018.11.010.

[26] C. Conejero-Goldberg, J.J. Gomar, T. Bobes-Bascaran, T.M. Hyde, J.E. Kleinman, M.M. Herman, S. Chen, P. Davies, T.E. Goldberg, APOE2 enhances neuroprotection against Alzheimer's disease through multiple molecular mechanisms, Mol. Psychiatry. 19 (2014) 1243–1250. https://doi.org/10.1038/mp.2013.194.

[27] V. Chouraki, S. Seshadri, Genetics of Alzheimer's Disease, in: Adv. Genet., 2014: pp. 245–294. https://doi.org/10.1016/B978-0-12-800149-3.00005-6.

[28] M.A. Rasmussen, B. Holst, Z. Tümer, M.G. Johnsen, S. Zhou, T.C. Stummann, P. Hyttel, C. Clausen, Transient p53 Suppression Increases Reprogramming of Human Fibroblasts without Affecting Apoptosis and DNA Damage, Stem Cell Reports. 3 (2014) 404–413. https://doi.org/10.1016/j.stemcr.2014.07.006.

[29] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics. 34 (2018) 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

[30] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools, Bioinformatics. 25 (2009) 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

[31] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics. 26 (2010) 841–842. https://doi.org/10.1093/bioinformatics/btq033.

[32] R Development Core Team, R: A language and environment for statistical computing., Vienna, Austria. (2017). https://doi.org/R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[33] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26. https://doi.org/10.1038/nbt.1754.

[34] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics. 30 (2014) 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

[35] D. Schenten, V.L. Gerlach, C. Guo, S. Velasco-Miguel, C.L. Hladik, C.L. White, E.C. Friedberg, K. Rajewsky, G. Esposito, DNA polymerase K deficiency does not affect somatic hypermutation in mice, Eur. J. Immunol. (2002). https://doi.org/10.1002/1521-4141(200211)32:11<3152::AID-IMMU3152>3.0.CO;2-2.

[36] F. Adikusuma, S. Piltz, M.A. Corbett, M. Turvey, S.R. McColl, K.J. Helbig, M.R. Beard, J. Hughes, R.T. Pomerantz, P.Q. Thomas, Large deletions induced by Cas9 cleavage, Nature. 560 (2018) E8–E9. https://doi.org/10.1038/s41586-018-0380-z.

Appendix A

Xdrop™ instrument, cartridges and reagents. Storing conditions and packaging (Parts) are also outlined.

Name	Cat. No.			
Xdrop™ Instrument	IN00100			
dPCR cartridge	CA10100			
dPCR gasket	GA10100			
dMDA cartridge	CA20100			
dMDA gasket	GA20100			
dMDA holder	HO10100			
Storage film	FI00100			
dMDA kit	RE20300	Part 1 (store at -20°C) dMDA mix (5x) dMDA enzyme	Part 2 (store at RT and 4□C) dMDA oil (RT) Break Solution (RT) Break colour (4□C)	
dPCR kit	RE10100	Part 1 (store at -20°C) dPCR mix (2x) Droplet dye	Part 2 (store at -20°C) dPCR buffer (2x)	Part 3 (store at RT and 4□C) dPCR oil (RT) Break Solution (RT) Break colour (4□C)
Cell sorter control kit*	CO10100	Part 1 (store at -20°C) Droplet dye dPCR buffer (2x)	Part 2 (store at 4°C) Control droplets	
Positive control primer kit*	CO10200	Part 1 (store at -20°C) dPCR control primers Enrichment validation primers Positive control DNA		
Primer test PCR kit*	RE10200	Part 1 (store at -20°C) dPCR mix (2X) qPCR dye (20x)		

*suggested items for workflow quality check.

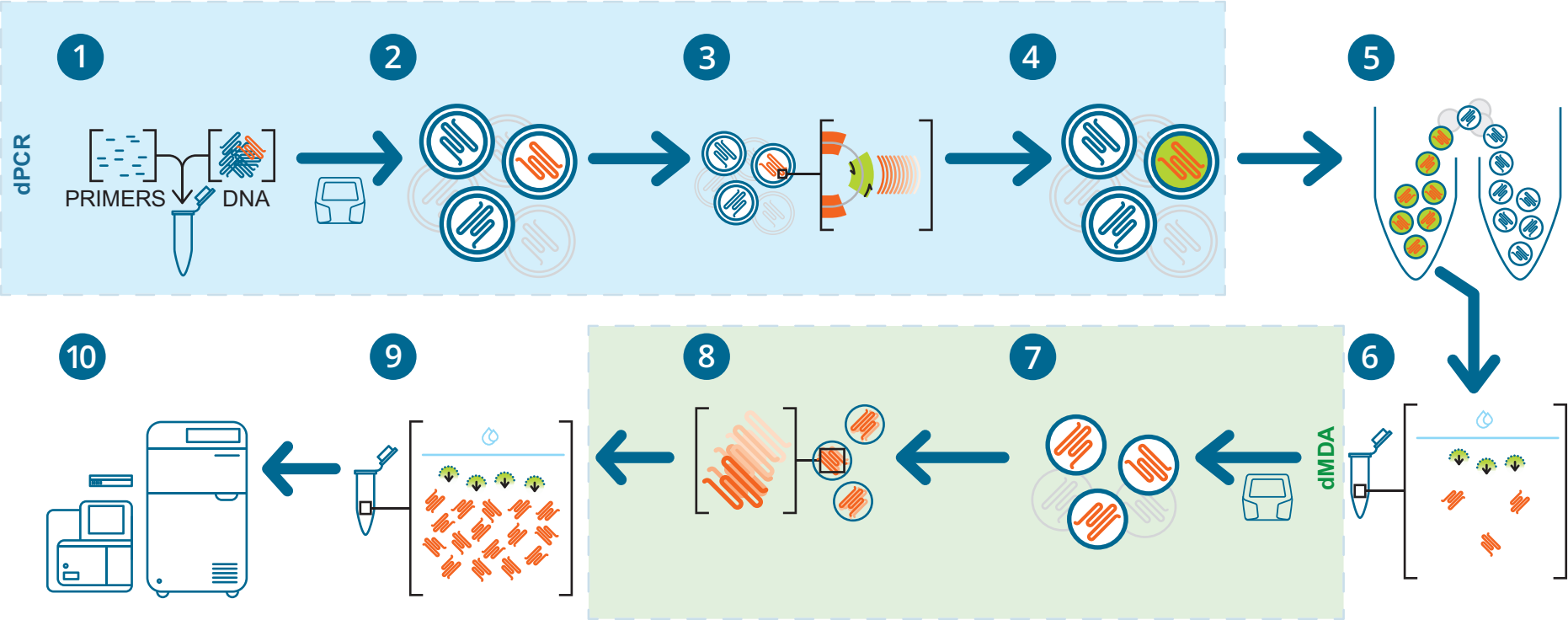
Appendix B

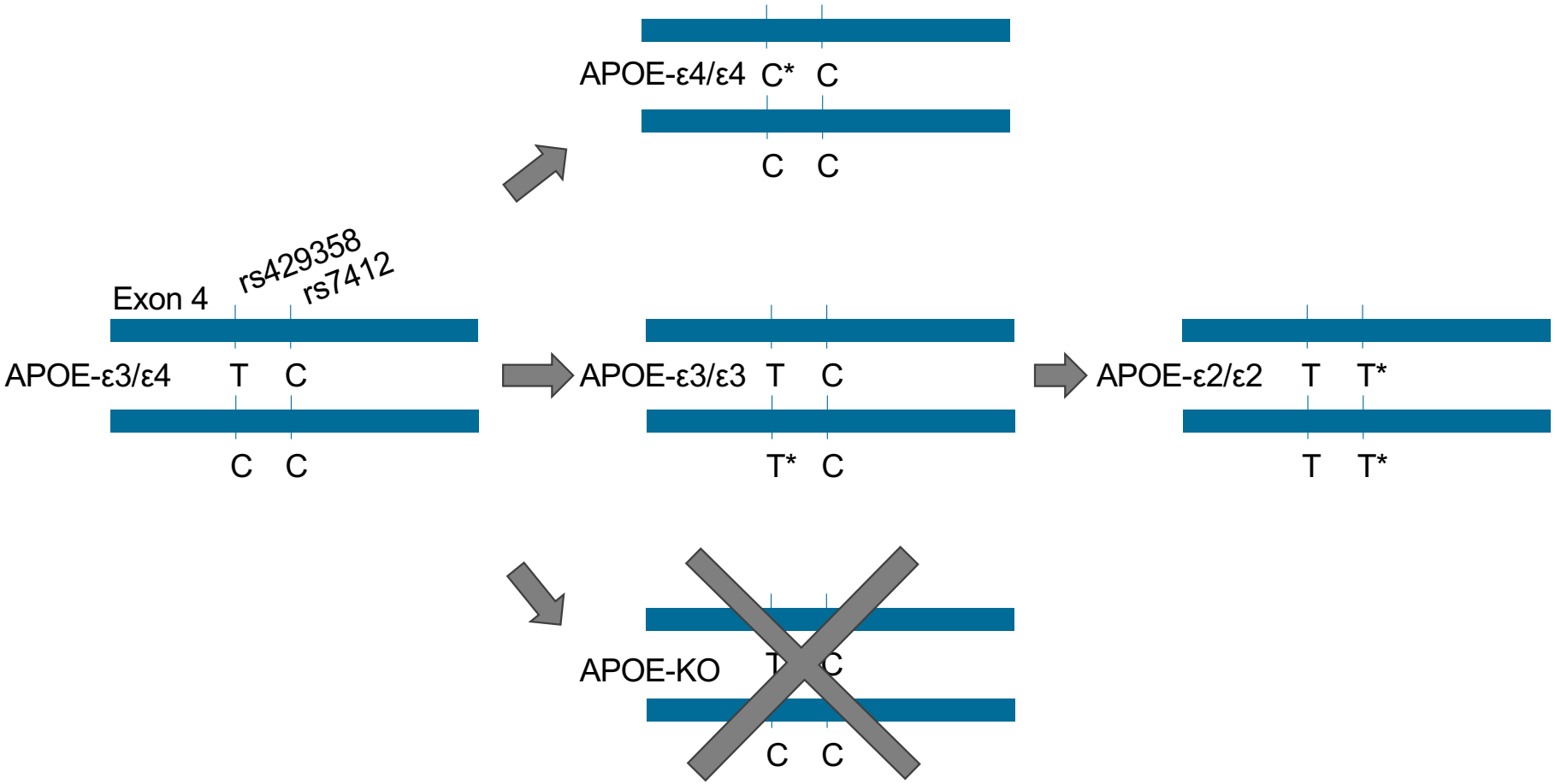
Aditional equipment, consumables and reagents needed.

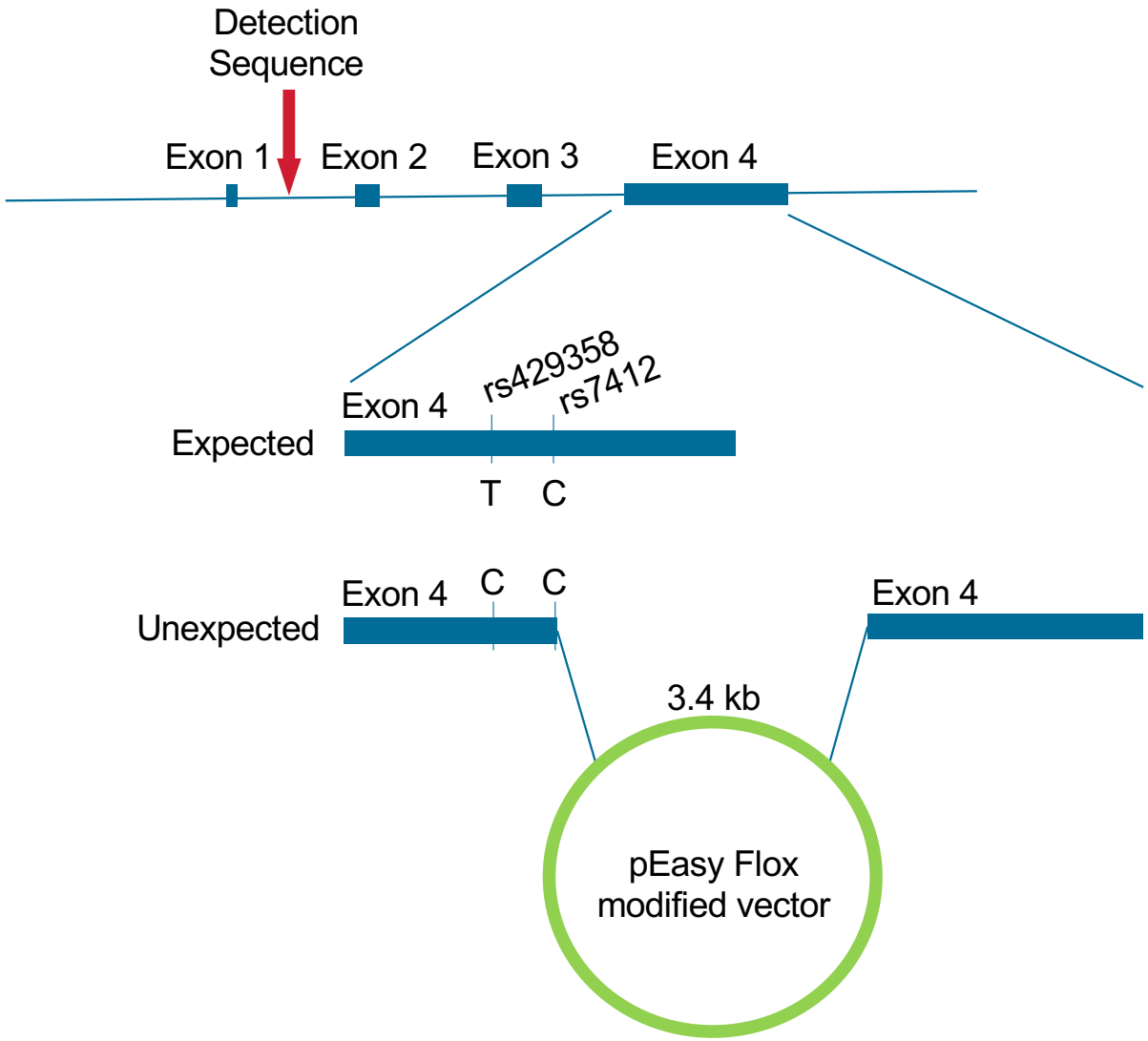
Equipment
Cell sorter instrument
Real-time PCR cycler
Thermal cycler with slow ramping function
Instrument for nucleic acid quantification e.g. Qubit™, Bioanalyzer™, TapeStation™, FEMTO Pulse™ or similar.
LAF (Laminar Air Flow) hood
Microcentrifuge
Vortex
Pipette set from P2 to P1000
Freezing blocks for both PCR tubes and microcentrifuge tubes

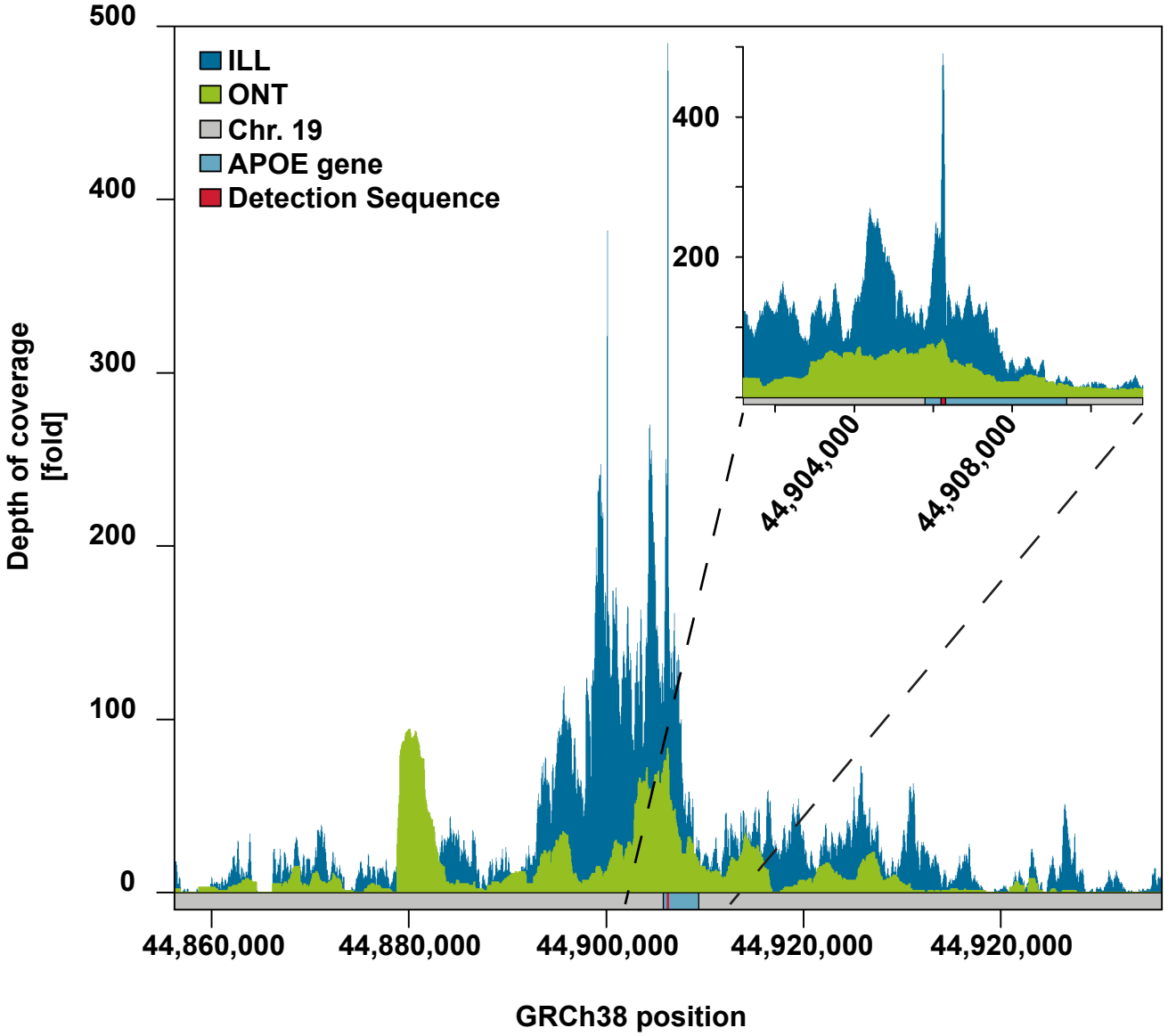
Consumables
Nuclease free PCR tubes and microcentrifuge tubes
Eppendorf DNA LoBind tubes 1,5 ml tubes (Catalog No. 0030108051)
Filtered pipette tips
Wide bore pipette tips for P200 pipette outer diameter of tip: 1 mm – 1,9 mm

Reagents
Nuclease-free water


dPCR
1
PRIMERS
DNA
2
3
4
5
6
dMDA
7
8
9
10


Exon 4
rs429358
rs7412
APOE-ε3/ε4
T
C
C
C
APOE-ε4/ε4
C*
C
C
C
APOE-ε3/ε3
T
C
T*
C
APOE-ε2/ε2
T
T*
T
T*
APOE-KO
T
C
C
C


Detection
Sequence
Exon 1
Exon 2
Exon 3
Exon 4
Exon 4
rs429358
rs7412
Expected
T
C
Exon 4
C
C
Unexpected
Exon 4
3.4 kb
pEasy Flox
modified vector


ILL
ONT
Chr. 19
APOE gene
Detection Sequence
Depth of coverage [fold]
500
400
300
200
100
0
44,860,000
44,880,000
44,900,000
44,920,000
44,920,000
GRCh38 position
400
200
44,904,000
44,908,000

chr19
378 bp
44,908,600 bp
44,908,700 bp
44,908,800 bp
44,908,900 bp
rs429358 (ref.T)
rs7412 (ref.C)
APOE
ε2/ε2
ONT
ε2/ε2
ILL
ε3/ε3
ILL
ε3/ε3
ONT