https://doi.org/10.15388/vu.thesis.248
https://orcid.org/0000-0002-1482-9536

VILNIUS UNIVERSITY

Žana
KAPUSTINA

The utility of modified nucleotides for high-throughput nucleic acid analysis

DOCTORAL DISSERTATION

Natural Sciences,
Biology (N 010)

VILNIUS 2021

This dissertation was prepared between 2016 and 2020 at the Institute of Biosciences, Life Sciences Center, Vilnius University and the Research and Development department at UAB Thermo Fisher Scientific Baltics. The research was supported by UAB Thermo Fisher Scientific Baltics and the scholarship from Vilnius University.

Academic supervisor – Prof. Dr. Arvydas Lubys (Vilnius University, Natural Sciences, Biology, N 010).

https://doi.org/10.15388/vu.thesis.248
https://orcid.org/0000-0002-1482-9536

VILNIAUS UNIVERSITETAS

Žana
KAPUSTINA

Modifikuotų nukleotidų taikymas plataus masto nukleorūgščių analizei

DAKTARO DISERTACIJA

Gamtos mokslai,
Biologija (N 010)

VILNIUS 2021

Disertacija rengta 2016– 2020 metais Vilniaus universiteto Gyvybės mokslų centro Biomokslų institute ir UAB „Thermo Fisher Scientific Baltics“ Mokslinių tyrimų ir eksperimentinės plėtros centre.
Mokslinius tyrimus rėmė UAB „Thermo Fisher Scientific Baltics“ ir Vilniaus universitetas.

Mokslinis vadovas – prof. dr. Arvydas Lubys (Vilniaus universitetas, gamtos mokslai, biologija, N 010).

TABLE OF CONTENTS

ABBREVIATIONS

ACN	acetonitrile
AMCA	7-amino-4-methylcoumarin-3-acetic acid
AMV	avian myeloblastosis virus
AUC	area under the curve
cDNA	complementary DNA
CuAAC	copper-catalyzed alkyne-azide 1,3-dipolar cycloaddition
ddN^{I}TP	5-iodo- or 7-iodo-7-deaza-2′,3′-dideoxynucleoside 5′-triphosphate
ddN^{N3}TP	5- or 7-deaza-7-(3-(2-azidoacetamido)prop-1-ynyl)-2′, 3′-dideoxynucleoside 5′-triphosphate
ddNTP	2′,3′-dideoxynucleoside 5′-triphosphate
DMSO	dimethyl sulfoxide
dN*TP	base-modified 2′-deoxynucleoside 5′-triphosphate
DNA-seq	DNA sequencing
dNMP	2′-deoxynucleoside 5′-monophosphate
dNTP	2′-deoxynucleoside 5′-triphosphate
dsDNA	double-stranded DNA
DTT	1,4-dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EMP	Earth Microbiome Project
exo-	exonuclease-deficient
FAM	fluorescein amidite
gDNA	genomic DNA
HIV	human immunodeficiency virus
HPLC	high performance liquid chromatography
HRP	horseradish peroxidase
IVT	*in vitro* transcription
LC-MS	liquid chromatography – mass spectrometry
M-MLV	Moloney murine leukemia virus
MNase-seq	micrococcal nuclease digestion with deep sequencing
MPE-seq	multiplexed primer extension sequencing
mRNA	messenger RNA
MTAS-seq	mRNA sequencing by terminator-assisted synthesis
N*TP	base-modified ribonucleoside 5′-triphosphate
NGS	next-generation sequencing
NTP	ribonucleoside 5′-triphosphate
ON	oligonucleotide
OTDDN or ddN^{ON}TP	oligonucleotide-tethered 2′,3′-dideoxynucleoside 5′-triphosphate

OTU	operational taxonomic unit
PAGE	polyacrylamide gel electrophoresis
PAP	poly(A) polymerase
PBS	phosphate-buffered saline
PCA	principal component analysis
PCR	polymerase chain reaction
PDB	Protein Data Bank
PEG	polyethylene glycol
PEX	primer extension
PPase	pyrophosphatase
PUP	poly(U) polymerase
R110	rhodamine 110 chloride
REG	carboxyrhodamine 6G
RNAP	RNA polymerase
RNA-seq	RNA sequencing
ROC	receiver operating characteristic
ROX	carboxy-X-rhodamine
rRNA	ribosomal RNA
RT	reverse transcriptase or reverse transcription
SAM	S-adenosyl-L-methionine
SBS	sequencing by synthesis
scRNA-seq	single-cell RNA sequencing
SELEX	systematic evolution of ligands by exponential enrichment
ssDNA	single-stranded DNA
st16S-seq	semi-targeted 16S rRNA gene sequencing
STAMP	single-cell transcriptomes attached to microparticles
TAMRA	carboxytetramethylrhodamine
TAP	*Thermoplasma acidophilum* inorganic pyrophosphatase
TBE	tris-borate-EDTA
TdT	terminal deoxynucleotidyl transferase
TEAAc	triethylammonium acetate
THPTA	tris-hydroxypropyltriazolylmethylamine
TSO	template switching oligonucleotide
UHRR	universal human reference RNA
UMI	unique molecular identifier
UTR	untranslated region
V1-V9	hypervariable regions of 16S rRNA gene
WGS	whole genome sequencing
wt	wild type

INTRODUCTION

While natural nucleotide modifications, such as methyl, hydroxymethyl, formyl, carboxy moieties, play important role in epigenetic regulation (Bilyard *et al*., 2020), the availability of synthetic modified nucleotides expanded the applicability repertoire of nucleic acids to many exciting fields, such as therapeutics, bioanalysis, chemical biology, catalysis, biosensing, and others (Dhuri *et al*., 2020; Xu *et al*., 2017; Lapa *et al*., 2016; Hollenstein, 2015; Hollenstein *et al*., 2008). Base modifications are usually introduced at the C5-position of pyrimidines or at the C7-position of 7-deazapurines (Jäger *et al*., 2005), and can range from small to bulky functional groups, including entities as large as proteins. Remarkably, natural DNA and RNA polymerases exhibit some plasticity in substrate recognition and can accept synthetic nucleotide analogs as substrates. This opens doors for the convenient enzymatic synthesis of functionalized nucleic acids.

Practical utility of modified nucleic acids in many cases relies on their biocompatibility. Although it was reported that phosphodiester, amide and triazole-based backbones are functional *in vitro* and even *in vivo* (Ciafrè *et al*., 1995; Kuwahara *et al*., 2009; El-Sagheer *et al*., 2011; Birts *et al*., 2014), template properties of modified DNA and RNA receive relatively less attention than studies of enzymatic catalysis with nucleotide analogs. Previous studies revealed that phosphate group itself is not essential for the ability to copy synthetic templates, and suggested CuAAC “click” reaction as good means to assemble biocompatible modified nucleic acids. High-fidelity replication through unnatural backbones would allow to unequivocally analyze the sequence of functionalized nucleic acids. It was observed that proofreading DNA polymerases stall at modification sites within template leading to the introduction of multibase deletions, while exonuclease-deficient enzymes are less likely to read through artificial backbones incorrectly (Shivalingam *et al*., 2017). Modified nucleic acids together with compatible enzymes make an attractive toolbox for the development of new molecular biology applications.

Chemoenzymatic approaches are paving their way to improve sample preparation for next-generation sequencing (NGS). “Click” reaction, termed chemical ligation, enables specific addition of alkyne-modified sequencing adapters to azido-modified chain terminators enzymatically introduced into sequenceable molecules. The technology proved to be feasible in various DNA and RNA sequencing workflows (Miura *et al*., 2018; Routh *et al*., 2015), offering easy generation of DNA and cDNA fragments, lower rates of chimera formation and the ability to prepare NGS libraries from ssDNA. Nonetheless, copper-mediated degradation of DNA, very low conversion efficiency and cumbersome protocols thus far impede wider adoption of such methods.

Aim and tasks

The **aim** of this study was to investigate the properties and utility of oligonucleotide-tethered 2′,3′-dideoxyribonucleoside 5′-triphosphates (OTDDNs, Fig. 1), with emphasis on labeling of DNA and cDNA molecules for high-throughput sequencing applications.

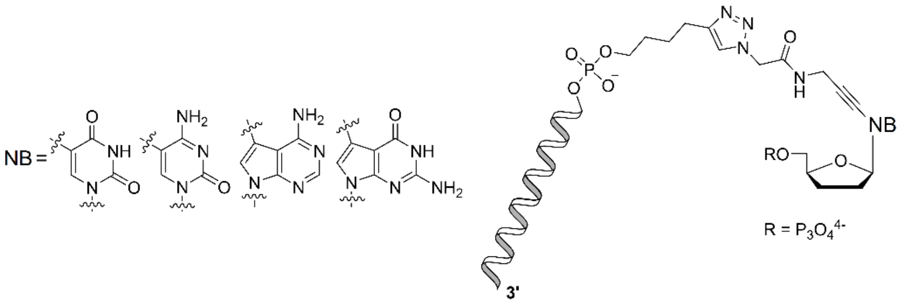


Figure 1. Chemical structure of oligonucleotide-tethered 2′,3′-dideoxyribonucleoside 5′-triphosphates. NB – nucleobase.

The following **tasks** have been defined to reach this aim:

1. To identify enzymes suitable for incorporation of OTDDNs into DNA or RNA.
2. To explore enzyme engineering possibilities in order to create new ways of synthesis of OTDDN-tagged nucleic acids.
3. To investigate template properties of OTDDN-containing nucleic acids and identify DNA polymerases able to read through unnatural triazole-based linkage.
4. To study the utility of OTDDNs for DNA and RNA sequencing.
5. To employ genomic DNA labeling by OTDDNs for the characterization of microbial communities.
6. To apply cDNA labeling by OTDDNs for high-throughput gene expression analysis.

Scientific novelty and practical value

In this work, oligonucleotide modification was used as a universal primer hybridization site to initiate the synthesis of a complementary strand for the amplification of labeled DNA fragments. Moreover, the use of dideoxynucleotides enabled termination of the nascent DNA or cDNA strand to obtain fragments of suitable length for short-read sequencing.

We demonstrated that efficient DNA labeling by OTDDNs can be achieved by Thermo Sequenase, CycleSeq, Sequenase V2.0 and TdT enzymes, cDNA can be labeled by Maxima, SuperScript IV, SuperScript II, RevertAid and HIV reverse transcriptases, and RNA – by poly(U) polymerase. Moreover, it was demonstrated that T7 RNA polymerase variant V783M, engineered towards the relaxed substrate discrimination, can synthesize chimeric dNMP-containing transcripts labeled by OTDDNs, which can be further directly used as PCR templates. Triazole-based linker within OTDDN is bypassed during the synthesis of complementary strand by Phusion exo-, Klenow fragment exo-, Thermo Sequenase and SuperScript IV DNA polymerases, with Phusion exo- exhibiting the best performance. Successful identification of enzymes for labeling and read-through enabled us to apply OTDDN technology for the preparation of fragment libraries for NGS. Importantly, it was demonstrated that oligonucleotide modification can contain regions of randomized sequence or affinity labels for convenient molecular barcoding or specific enrichment of tagged molecules, respectively.

Stochastic nature of OTDDN incorporation and simultaneous tagging of nascent strand with the universal oligonucleotide laid the foundation for the development of *semi-targeted* sequencing approach. It was demonstrated that labeling of random primer extension products allows to prepare whole genome or whole transcriptome libraries, while the extension of specific primers opens doors to study *a priori* unknown sequence regions nearby defined target loci. This strategy proved to be useful for the analysis of microbial communities. A new method, termed semi-targeted 16S rRNA gene sequencing (st16S-seq), was developed in this work to capture regions of bacterial genomes upstream of 16S rRNA gene. This technique offers substantial advantages over conventional approaches, including precise determination of 16S rRNA gene copy numbers, better characterization accuracy and less prominent dependency of taxon capture efficiency on primer design.

In RNA sequencing field, a new method, termed mRNA sequencing by terminator-assisted synthesis (MTAS-seq), was developed to simplify gene expression analysis workflows for both bulk and single-cell analyses. It was demonstrated that MTAS-seq streamlines protocols, enables easy molecular barcoding of cDNA fragments, is compatible with direct library preparation from cell lysates and generates data of equivalent or better quality than conventional techniques. Moreover, 3′UTR enrichment via oligo(dT) primed reverse transcription enables the analysis of alternative polyadenylation patterns in eukaryotic transcriptomes. Interestingly, the use of OTDDN with oligonucleotide modification corresponding to the full-length sequencing adapter enabled to sequence cDNA libraries without PCR amplification, which was not previously possible on Illumina™ DNA sequencing machines.

This work suggests OTDDNs as a promising tool for NGS applications, offering more convenient protocols and opportunity to develop new library preparation methods. Techniques developed in this study are covered by international patent application.

Statements to be defended

1. OTDDNs may act as substrates for DNA and RNA polymerases with natural or engineered ability to incorporate nucleotide analogs.
2. OTDDN-labeled DNA can serve as a template for the synthesis of complementary strand. Exonuclease-deficient DNA polymerases can copy through triazole-based linker.
3. Primer extension and OTDDN incorporation by Thermo Sequenase or SuperScript IV RT and subsequent synthesis of a complementary strand starting from the primer hybridization site within OTDDN enable DNA and RNA sequencing applications. Template coverage depends on the sequence of primers used in labeling reaction.
4. Semi-targeted sequencing of bacterial 16S rRNA gene improves high-throughput characterization of microbial communities by providing precise information about 16S rRNA gene copy numbers and improving classification accuracy at species level.
5. OTDDN labeling improves gene expression analysis workflows by providing single-tube protocol, easy molecular barcoding of cDNAs via randomized sequences embedded within OTDDNs and offering compatibility with direct reverse transcription from whole cell lysates.

1. LITERATURE OVERVIEW

Naturally occurring non-canonical nucleotides enrich the chemical diversity of DNA and RNA and enable them to effectively execute complex functions in a cell. The position of the modified nucleotides and their chemical structure establishes a second layer of genetic information which defines the research areas of epigenetics. While the repertoire of nucleotide modifications found in RNA is remarkably wide, the number of modifications found in DNA is small. This is related to the crucial but limited function of DNA as a carrier of genetic information (Carell *et al.*, 2012).

The interest in non-canonical nucleotides and functionalized DNA and RNA polymers has been growing for the past decades (Nakatani & Tor, 2016), with rapidly expanding applications in chemical biology, bioanalysis, therapeutics, as illustrated by the development of molecular reporters (Xu *et al.*, 2017), aptamers (Lapa *et al.*, 2016; Zhou & Rossi, 2017), catalytic nucleic acids (Hollenstein, 2015), biosensors (Hollenstein *et al.*, 2008), artificial biomolecules used in programmable coding of a certain function *in vivo* (Tarashima *et al.*, 2016), and many more.

Chemical alterations of nucleic acids include the diversification of nucleobase, sugar moiety or phosphodiester backbone, and various combinations of these modifications (Fig. 1.1). To unlock the full potential of increasing number of nucleotide analogs, the availability of polymerases with an expanded range of acceptable substrates is necessary. Interestingly, natural polymerases exhibit some plasticity in substrate recognition – modifications at certain positions are readily tolerated, moreover, even the creation and replication of unnatural base pairs are possible, which allows to expand the genetic code (Hoshika *et al.*, 2019; Marx & Betz, 2020).

This overview will focus on diverse opportunities offered by implementation of modified nucleic acids as synthetic biomolecular blocks. Furthermore, the progress made in understanding of structural prerequisites for successful catalysis of incorporation of modified nucleotides by polymerases will be reviewed. The spectrum of nucleic acid modifications will be confined to base alterations which is the subject of the present dissertation.

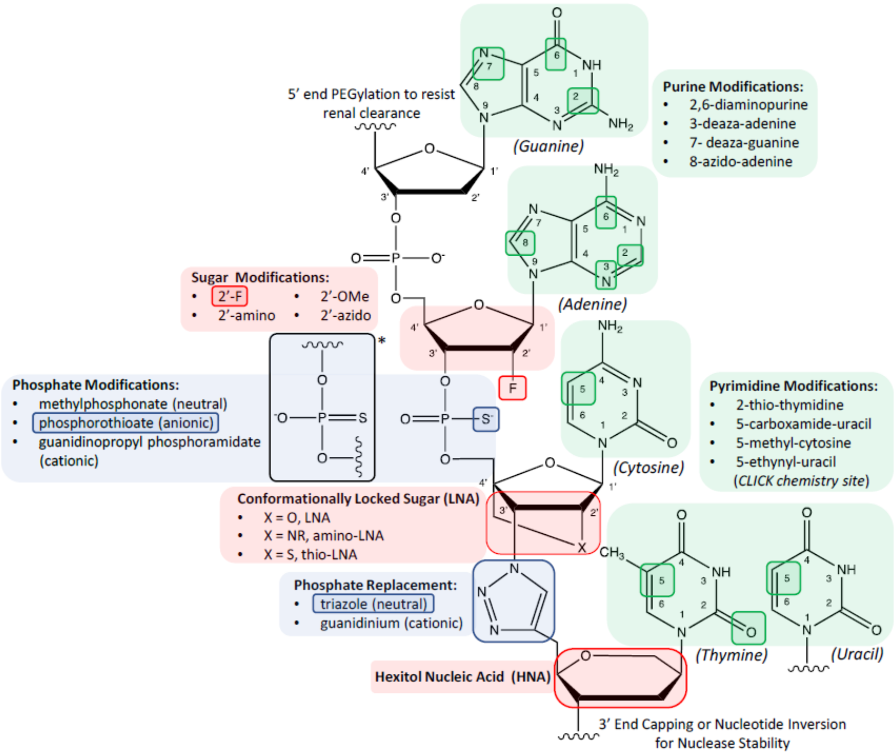


Figure 1.1. An illustration of various chemical modifications of DNA structure ranging from simple site-specific atomic substitutions to more complex molecular replacements bearing little resemblance to the natural structure (by Ochoa & Milam, 2020).

1.1 Base-functionalized nucleic acids: a versatile toolbox

For successful utilization, the design of functionalized nucleotides should comply with the following conditions, originally defined for the process of systematic evolution of ligands by exponential enrichment (SELEX; Perrin *et al.*, 1999), however applicable for other molecular biology techniques:

- should not interfere with the base pairing (Watson-Crick and Hoogsteen);
- nucleotides should be substrates of the corresponding DNA or RNA polymerases;
- the incorporation of a modified nucleotide should be efficient at any position or context of the sequence;
- in some instances, the functionalized sequence should be a template for the corresponding polymerases.

Base modifications are usually attached at the C5-position of pyrimidines or at the C7-position of 7-deazapurines because such substituents are well accommodated in the major groove of DNA without disturbing the helical structures (Jäger *et al.*, 2005). Purines modified at C8-position were also reported, however those with bulky groups appeared to be poor substrates for polymerases (Cahová *et al.*, 2008). Interestingly, N^4-acylated 2′-deoxycytidines were found to be efficiently incorporated by a variety of polymerases, even when N^4-substituents were sterically demanding, e.g. benzoylbenzoyl moiety (Jakubovska *et al.*, 2018).

The classical synthesis approach for base-modified precursors consists of multistep synthesis of aminopropargyl-, aminopropenyl- or aminopropyl-substituted dNTPs or NTPs through palladium-catalyzed cross-coupling reaction (Shaughnessy & DeVasher, 2005; Sonogashira *et al.*, 1975) of halogenated nucleosides with CF_3CO-protected modifications, followed by triphosphorylation and deprotection. The desired functional group is then attached to the amino group via amide bond formation. Such workflow is rather laborious and produces low yields of the desired product, as may be exemplified by a 5-step synthesis of amidine-modified dUTP with overall yield of 8% (Jäger *et al.*, 2005).

Palladium-catalyzed cross-coupling reactions are widely used for C-C bond formation, however they were hardly adaptable for nucleotides, since reaction media usually led to poor solubility of nucleotide precursors. With the discovery of water-soluble catalytic systems, aqueous-phase cross-coupling reactions have been developed, allowing a straightforward introduction of unprotected functional groups directly into dNTPs or NTPs (Fig. 1.2). The Sonogashira cross-coupling reactions of 5-iodo-dUTP with fluorescein-linked terminal acetylenes demonstrated by Thoresen *et al.* were the very first cross-coupling reactions with iodo-modified dNTPs (Thoresen *et al.*, 2003). The required halogenated nucleotides are accessible by chemical triphosphorylation of halogenated nucleosides (5-I-dU, 5-I-dC, 7-I-7-deaza-dA and 7-I-7-deaza-dG; Kovacs & Ötvös, 1988).

Most base-modified nucleotides reported in literature are C5-substituted uracil derivatives, whereas the 7-deazaguanines are only scarcely reported due to difficult multistep synthesis process of the nucleoside intermediates (Hocek, 2014).

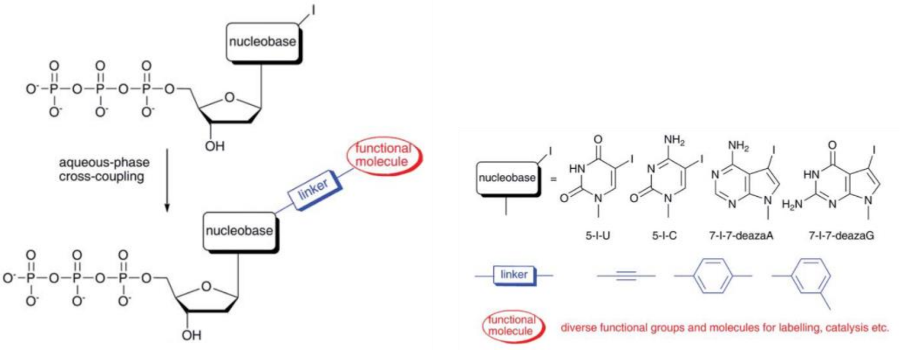


Figure 1.2. General scheme for the construction of base-functionalized nucleotide analogs (by Hocek & Fojta, 2008).

Cycloaddition reactions are one of the most useful and popular approaches for conjugation of functional groups, in particular copper-catalyzed alkyne-azide 1,3-dipolar cycloaddition (CuAAC), or “click“ reaction. “Click“ chemistry has been extensively used for bioconjugations and modifications of DNA because of its bioorthogonality and efficiency. For compatibility, nucleobases within DNA should be modified either by terminal alkyne or by azido groups. The alkynes are compatible both with phosphoramidite synthesis and polymerase incorporations, and a number of alkyne-modified phosphoramidites or dNTPs were used for synthesis of alkyne-labeled DNA with subsequent CuAAC modifications in the major groove. Azido group is not compatible with phosphoramidite synthesis on solid support, but it can be still introduced into DNA by enzymatic incorporation of azido-modified nucleotides (Ivancová *et al.*, 2019; Panattoni *et al.*, 2018). It is worth noting that typically DNA labeling is a postsynthetic process, i.e. the substrate for CuAAC modification is dsDNA, ssDNA or oligonucleotides (ONs) containing one, several or many modified nucleotides introduced by various enzymatic methods, such as primer extension (PEX) or PCR.

1.1.1 Dye-labeled nucleotides

Fluorescent nucleobases have emerged as an extraordinary tool for the molecular-level understanding of nucleic acid structure, function, locations and interactions. Aromatic heterocycles within purines and pyrimidines are receptive to diverse modifications, and even minimal structural and electronic perturbations can dramatically transform their photophysical properties. Moreover, the boundaries of natural molecular skeletons of purines and pyrimidines can be broken further by complete substitution of natural nucleobases by fluorescent non-canonical analogs.

One strategy to produce chromophoric base analogs is replacing the natural nucleobases with known fluorophores, typically polycyclic aromatic hydrocarbons. Such compounds constitute a non-canonical family of chromophoric base analogs that are unable to form Watson-Crick hydrogen bonds, however, can be used to investigate enzyme-substrate recognition mechanisms. Pyrene nucleotide was employed to demonstrate that nucleotides should be of correct size and shape to fit the enzyme active site against a template base, indicating the importance of steric complementarity in the fidelity of DNA synthesis (Matray & Kool, 1999). Advances in bioorganic chemistry field led to the introduction of hundreds of fluorescent nucleobases, all of which exhibit distinct base pairing and stacking abilities or emission profiles (Sinkeldam *et al.*, 2010).

Among the base-modified fluorescent nucleotides, extended nucleotides carrying chromophores tethered to the nucleobase with or without a linker are the most frequently reported due to their broad structural variety and retained ability to form Watson-Crick hydrogen bonds. In studies conducted by Wagenknecht and co-workers, “click“ chemistry was employed to postsynthetically conjugate ethynyl-modified fluorophore Nile Red to the C5-position of uridine after incorporation of 5-iodo-2'-deoxyuridine into ON and *in situ* formation of the intermediate azide. Researchers aimed to investigate aspects of ON structure and effects of chromophore stacking using an acyclic linker scaffold (Beyer & Wagenknecht, 2010; Lachmann *et al*., 2010). Another example is the work by Østergaard *et al*. who attached a pyrene residue to the C5-position of uridine through a triazole moiety. The resulting compound exhibited favorable mismatch discrimination by significantly decreasing its quantum yield in the presence of mismatched sequence – this feature proved to be useful for the detection of single nucleotide polymorphisms using fluorescence (Østergaard *et al.*, 2010). Hocek and co-workers developed modified 2′-deoxycytidines bearing environment-sensitive fluorophores and designed DNA probes for sensing protein-DNA interactions. After the conversion of nucleoside to its triphosphate analog, fluorophores including dimethylaminobenzylidene cyanoacetamide or tryptophan-based imidazolinone were incorporated into DNA probes within modified dCTPs and showed light-up response upon binding to p53 or single-strand binding protein (Hocek, 2019).

Seela *et al.* have synthesized various 7-deaza and 8-aza-7-deazanucleosides related to 2′-deoxyadenosine and 2′-deoxyguanosine containing 7-octadiynyl or 7-tris(propargylamine) pendant groups and the corresponding ONs containing these modifications. The authors afterwards chemically ligated fluorophores, such as 9-azidomethylanthracene, 3-azido-7-

hydroxycoumarin or 1-azidomethylpyrene using “click“ reaction to obtain fluorescently labeled ONs (Seela & Pujari, 2010; Ingale *et al.*, 2012).

In the field of nucleic acid analysis, sequencing technologies – undoubtedly revolutionary – contribute substantially to the development of biological sciences for several decades now. Modern version of Sanger sequencing (Sanger *et al.*, 1977), as well as certain next-generation sequencing (NGS) platforms, e.g. sequencing by synthesis (SBS) commercialized by Illumina™, rely on fluorescence detection.

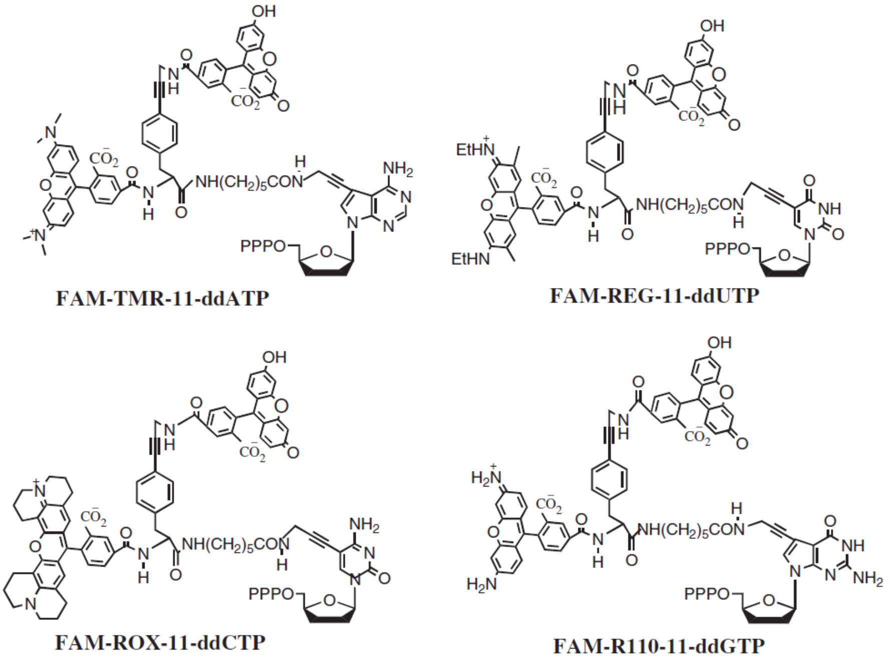


Figure 1.3. A four-color set of energy-transfer dye-labeled terminators (by Kumar & Fuller, 2007).

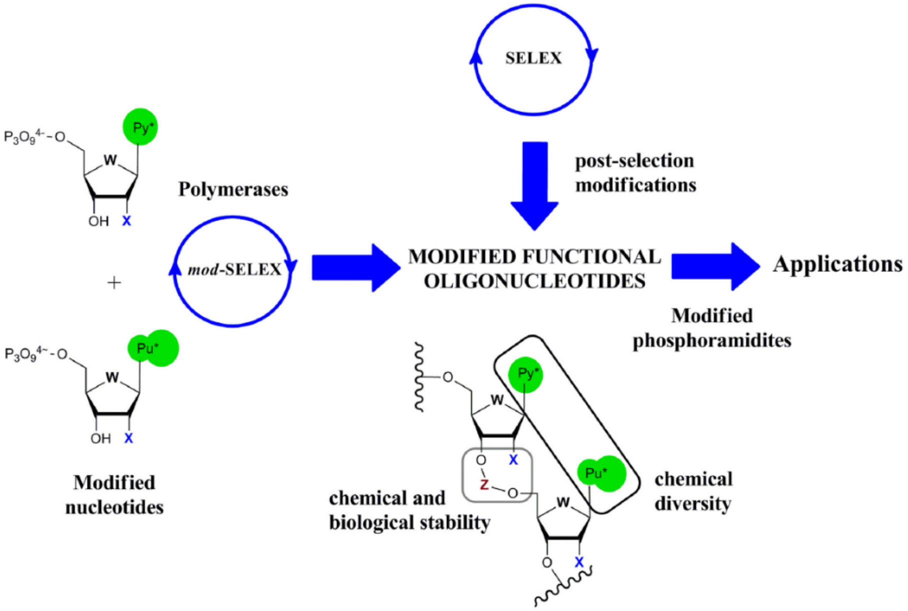
Automated Sanger DNA sequencing can be performed in two ways: employing dye-labeled primer, where fluorescent dyes are conjugated to the 5′ terminus of the primer ON or using dye-labeled chain terminators (dideoxynucleoside triphosphates). The latter option is more convenient since a single primer extension reaction is required per template (Rosenblum *et al.*, 1997). The rhodamine dyes (R110, REG, TAMRA and ROX) are widely adopted for DNA sequencing as these dyes absorb and emit light optimally at different wavelengths, however their quantum yields are different. To compensate for these differences, energy transfer principle can be employed by attaching an additional donor chromophore at a certain distance from an acceptor chromophore. An example of energy-transfer dye-labeled dideoxynucleoside triphosphates containing fluorescein (FAM) donor dye and an acceptor rhodamine dye modification is given in Fig. 1.3. These modified terminators proved to be good substrates for Thermo Sequenase DNA polymerase (Kumar & Fuller, 2007).

Ultra-high-throughput SBS has become possible by exploiting reversible terminators – nucleotides bearing fluorescent dye attached via a cleavable linker to the nucleobase and having a 3′-OH group capped with a small chemically reversible moiety (3′-*O*-azidomethyl). It was shown that these nucleotide analogs can be efficiently incorporated during a solution-phase DNA extension reaction, and both modifications can be subsequently removed in aqueous solution by tris(2-carboxyethyl)phosphine treatment. Various alternatives of sequencing by synthesis chemistry were reported, including the use of two types of modified nucleotides – cleavable 3′-*O*-azidomethyl-blocked dNTP reversible terminators together with fluorescently tagged ddNTP irreversible terminators (Ju *et al.*, 2006; Guo *et al.*, 2008; Knapp *et al*., 2011).

1.1.2 Functional oligonucleotides

Although some examples of functional oligonucleotide activities exist in nature, as in the case of ribozymes, microRNAs or riboswitches, a set of catalytic (ribozymes and DNAzymes) and molecular recognition (aptamers) oligonucleotides have been synthetically prepared, since they can be obtained using *in vitro* evolution techniques. This methodology was simultaneously developed by three independent groups in 1990s (Ellington & Szostak, 1992; Robertson & Joyce, 1990; Tuerk & Gold, 1990) and was named SELEX. Eventually, many different variations of the technique were developed to achieve better selectivity, binding constants and simpler experimental protocols.

Historically, proteins have dominated the pool of available catalysts and affinity reagents, likely because of the diverse array of available amino acid side chains. In contrast, DNA and RNA comprise of only four nucleotide building blocks, and they all possess a relatively similar repertoire of functional groups. Moreover, wild type based nucleic acid biocatalysts and aptamers have limited tolerance to nucleases and might also be degraded chemically. To overcome these limitations and augment the structural diversity, functional oligonucleotides with various structural chemical modifications can be obtained either via modified-SELEX (*mod*-SELEX) or through post-selection modification (Fig. 1.4; Dellafiore *et al*., 2016; Meek *et* *al*., 2016). A number of research groups have succeeded in generation of functional oligonucleotides with base-modified nucleotides.



SELEX
post-selection modifications
$P_3O_9^{4-}$-O
Py*
W
OH X
Polymerases
mod-SELEX
+
MODIFIED FUNCTIONAL OLIGONUCLEOTIDES
Applications
Modified phosphoramidites
Pu*
Modified nucleotides
O
Z
chemical and biological stability
chemical diversity

Figure 1.4. Schematic illustration of the alternative routes to obtain modified functional oligonucleotides. Py* - modified pyrimidines; Pu* - modified purines; X – F, OMe, NH_2; W – O, S; Z – phosphate, phosphorothioates, boronate esters (by Dellafiore *et al*., 2016).

An early example of successful application of *mod*-SELEX was reported by Latham and co-workers who selected an aptamer for thrombin from a library in which all thymines were replaced by 5-pentynyl-dU. This aptamer showed unique secondary and tertiary structures as compared to previously selected wild type DNA aptamer (Latham *et al*., 1994). Jensen *et al*. generated a base-modified RNA aptamer specific to the HIV REV protein from the library of modified RNA containing 5-iodouridine. The aptamer demonstrated higher binding affinity than its natural counterpart and was able to form a crosslink with the target protein upon UV irradiation (Jensen *et al*., 1995). Sawai and colleagues reported an example of *in vitro* selection of DNA aptamer for thalidomide from a library containing dUTP analog decorated with a cationic amine attached via a hexamethylene linker arm to the C5-position of the nucleobase. Researchers showed that the most proficient aptamer crucially depended on the presence of modifications since the corresponding natural DNA sequence lost all binding-propensity (Sawai *et al*., 2001; Shoji *et al*., 2007). While first fruitful examples established precedence for the idea of using unnatural nucleotides in SELEX, the true potential of such approach was demonstrated by SomaLogic, Inc. who executed selection from oligonucleotide libraries containing dU/C bearing a hydrophobic functional group (e.g. benzyl, tryptamino) at the C5-position of the nucleobase. SELEX was used to screen libraries of such oligonucleotides equipped with amino acid mimics to identify slow off-rate aptamers called SOMAmers. SomaLogic team has established multiple SOMAmer based proteome assays with over 3000 different protein targets identified so far (Gold *et al*., 2010; Ochoa & Milam, 2020).

To allow the access to a larger palette of bulkier functional groups and avoid enzymatic incompatibility that often restricts the use of such modifications, SELEX variants with post-selection modification were developed, including click-SELEX (Pfeiffer *et al*., 2018). In this method, all dT nucleotides in the library are replaced with C5-ethynyl-2′-deoxyuridine which introduces multiple alkyne functional groups into the sequences. The library is then functionalized via a “click“ reaction with an azide, which in the case of initial demonstration was 3-(2-azidoethyl)indole. Tolle *et al.* used this approach to select an aptamer for the cycle-3 GFP. In addition, a series of azides were utilized to evaluate the importance of the specific indole moiety used for the selection process. Aptamer variants functionalized with the alternative azides were not capable of binding the target protein even when the azide was structurally similar to the indole used during the selection process (Tolle *et al*., 2015). This highlights the importance of the appended functional groups in target binding.

A very similar strategy, termed SELMA (selection with modified aptamers), was applied to generate DNA scaffolds with C5-ethynyl-2′-deoxyuridine moieties that were further glycosylated using glycan azides. After selection of the most antigenic clusterings of glycan, enriched sequences were amplified and reglycosylated to be used in next selection steps (Horiya *et al.,* 2014).

The versatility with which side chains having a wide variety of sizes and chemical properties can be conveniently introduced using SELEX with post-selection modification is anticipated to greatly accelerate the discovery of nucleic acid catalysts and affinity agents possessing novel functions (Meek *et al*., 2016).

1.1.3 Bulky modifications

Large functional groups are often attached to the nucleobases via linkers that vary in composition, length and flexibility. Employing this strategy, a variety of even very bulky modifications can be incorporated into DNA.

Along these lines, it has been shown that some DNA polymerases are capable of efficient incorporation of nucleotides modified with entities as large as proteins.

Sørensen and co-workers demonstrated that TdT can accept nucleoside triphosphates tethered to large biomolecules as substrates and incorporate such conjugates to the 3′ terminus of any native ON. Five different macromolecules that belong to different structural classes were tested as “cargos“ on the nucleobases of dNTPs: cyclic integrin targeting peptide, two polyethylene glycol polymers of different length, G3.5 PAMAM dendrimer and streptavidin. Labeling of oligonucleotides with functionalized dNTPs was highly efficient – all reactions achieved 99% conversion or better except for streptavidin which yielded 93% of streptavidin-labeled product (Sørensen *et al.*, 2013).

Welter and colleagues succeeded in site-specific incorporation of approximately 40 kDa glycoprotein horseradish peroxidase (HRP) from *Amoracia rusticana* into a nascent DNA strand without compromising the ability of the enzyme to produce colorimetric signals through the oxidation of dye substrates. Horseradish peroxidase activated with maleimide group through the conversion of the lysine residues with maleimidocaproic acid N-hydroxysuccinimide ester were tethered to two variants of thymidine analogs bearing ω-mercaptocarboxylic acid-based linkers of different lengths at the C5-position (dT^{7SH}TP and dT^{15SH}TP). Conjugation was performed through thiol-maleimide reaction (Fig. 1.5).

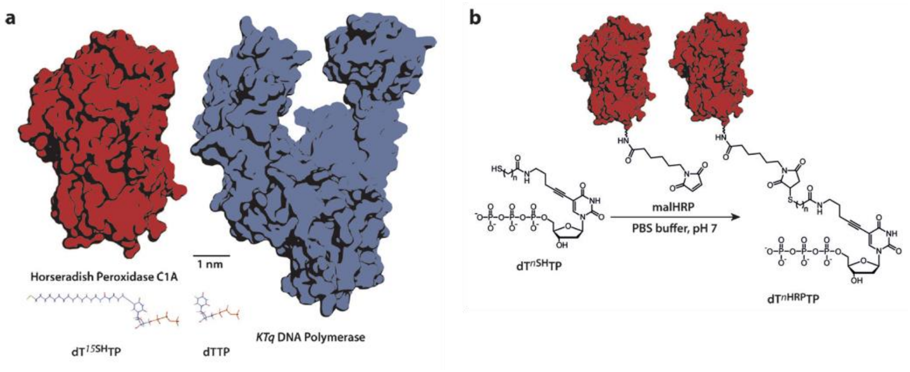


Figure 1.5. Conjugation of horseradish peroxidase to the nucleobase of dTTP. **(A)** – size comparison between HRP C1A, *KTq* DNA polymerase and a modified dTTP bearing a C_{15} thiol linker (dT^{15SH}TP). **(B)** – coupling of dT^{nSH}TP with malHRP (by Welter *et al.*, 2016).

Enzyme-labeled nucleotides, despite being more than 100-fold larger than the natural substrates, were accepted by the KlenTaq DNA polymerase. Modified nucleotide having longer C_{15} linker appeared to be a better substrate than the one with the shorter linker. Even multiple incorporations of HRP-modified $dT^{15SH}TP$ but not $dT^{7SH}TP$ were observed indicating that too short distance between the nucleotide and its bulky modification interferes with DNA polymerase activity. The authors used template-dependent protein-modified nucleotide incorporation to create site-specific DNA-protein conjugates with the possibility of naked-eye detection of the presence or absence of target sequence (Welter *et al.*, 2016).

Later, the authors from the same research group reported the development of antibody-modified nucleotides whose modification was larger than the DNA polymerase employed for incorporation. Nucleotide analogs $dC^{pAb}TP$ were incorporated into growing DNA strand in a sequence-specific manner by KOD DNA polymerase. The recognition of the antibody was not abolished by the conjugation – it was recognized by a secondary antibody bearing a signal-generating HRP enzyme thus enabling a colorimetric read-out of nucleotide incorporation (Balintová *et al.*, 2018).

Naked-eye detection of DNA with single-base resolution was also achieved through the use of oligonucleotide-modified nucleotides. ON-conjugated dUTP was accepted by Klenow Fragment (exo-) and KlenTaq DNA polymerases and successfully participated in PEX. Subsequently, ON modification served as a primer to initiate rolling circle amplification in the presence of complementary circular template. Extended single stranded product then captured 3′-flanking region of a G-quadruplex DNAzyme sequence which possesses HRP-like activity. Colorimetric signal was generated via hemin-mediated oxidation of 2,2′-azino-bis(3-ethyl-benzothiazoline-6-sulfonate) ($ABTS^{2-}$) by H_2O_2 (Verga *et al.*, 2016).

The opportunity to incorporate nucleotides that are covalently attached to large biomolecules, even whole proteins, is highly promising for future applications, such as high-performance diagnostic assays, fundamental research on molecular recognition, synthesis of DNA nanostructures, and others (Niemeyer, 2010).

1.1.4 (De)stabilizing effects of modified nucleotides

High potential of oligonucleotides as agents in diagnostic and therapeutic applications is hinged on high affinity and specificity of Watson-Crick hybridization. Nevertheless, much research has been devoted to the discovery of modifications which would improve the biostability of ONs while maintaining the hybridization characteristics of natural DNA.

Buhr and colleagues described the properties and antisense activities of ONs containing 7-(1-propynyl)-7-deaza-2′-deoxyguanosine and 7-(1-propynyl)-7-deaza-2′-deoxyadenosine. Researchers showed that 7-propynyl analogs bind to target RNA with higher affinity than natural purines. The increase in binding affinity is likely due to increased stacking interactions leading to more favorable enthalpy of binding. Interestingly, 7-propynyl-dG did enhance the antisense activity of ONs, however 7-propynyl-dA decreased the activity. The possible reason for such dichotomy may be the influence of adjacent sequence context or different ability to recruit RNase H cleavage of the complementary RNA (Buhr *et al*., 1996).

The properties of more than 200 modifications within ONs were evaluated by Freier and co-workers. In agreement with abovementioned study, the authors identified 7-propynyl-7-deazapurines as well as 7-halo-7-deazapurines as the most stabilizing purine base modifications and explained this effect by increased stacking of the modified purine rings. Experiments with thymine showed that substitution of the 5-methyl group with a halogen had little effect and substitution with a methoxy-ethoxy-methyl group was destabilizing. The highest stabilizing effect was obtained with 5-propynyl dU. Analogs containing amino-ethyl-3-acrylimido modifications at the 5-position showed some stabilizing effect most probably due to acrylimido group contribution to stacking, similarly to the propyne substitution. Slight positive effect on duplex stability was also observed for 5-amino-hexyl-substituted pyrimidines and was attributed to shielding of the negative phosphate charges in unmodified hybrid duplexes (Freier & Altmann, 1997).

Seela and colleague reported the synthesis and enzymatic incorporation of pyrazolo[3,4-*d*]pyrimidine 2′-deoxyribonucleoside triphosphates, of which the 7-bromo derivative harmonized the stability of DNA duplexes, i.e. the stability was no longer dependent on the base pair composition as dA•dT base pair became as stable as dG•dC without compromising sequence specificity (Seela & Becher, 2001). The same group also investigated the effects of unpaired terminal nucleotides (dangling ends) on thermal stability of DNA duplexes. They reported that the incorporation of a highly polarizable nucleotide residue (e.g. 8-aza-7-deaza-2′-deoxyisoguanosine or its 7-bromo or 7-iodo-substituted derivatives) into the inner part of a parallel or an antiparallel ON with a random sequence can enhance duplex stability if the ability to form Watson-Crick hydrogen bonds is retained (Rosemeyer & Seela, 2002).

SomaLogic, Inc. team systemically evaluated stabilizing effects of various modifications in the context of thermodynamic properties of 5-N-carboxamide modified SOMAmers (see chapter 1.1.2). The results indicated that bulky hydrophobic modifications, such as 5-[N-(1-naphthylmethyl)-carboxamide]-2′-deoxyuridine or 5-[N-(2-naphthylmethyl)carboxamide]-2′-deoxyuridine, were destabilizing, while the most stabilizing modification was an aliphatic isobutyl moiety. The hydrophilic modifications were also stabilizing, however the overall effect was modest. Smaller aromatic groups had neutral or mildly positive effect. Changes in thermal stability of the hybrid duplexes relative to DNA were explained by offsetting effects of enthalpy and entropy, with the enthalpic effect being dominant (Wolk *et al.*, 2015).

In some instances, it is desirable to exploit modifications to *destabilize* DNA duplex, as in the case of caged ONs whose activity is triggered by photolabile protecting groups. Upon irradiation with light, the protecting group is removed, restoring the activity of ON with high spatial and temporal control. Seyfried *et al.* investigated the destabilization of duplexes by single photolabile protecting groups attached at the Watson-Crick site of nucleobases. Researchers showed that diphenylmethyltriazole-coumarin conjugated in (*S*)-configuration to the nucleobase exhibits the largest duplex destabilization (ΔT_m=15.8°C) ever measured for a single base-caged DNA (Seyfried *et al.*, 2018).

1.2 Enzymatic processing of modified nucleic acids

The wide variety of modified nucleotides and the plethora of commercially available polymerases are two powerful tools in the molecular biologists′ workshop. Enzymatic routes to obtain modified nucleic acids are flexible, can produce single- or double-stranded products, and have fewer restrictions on product length (Whitfield *et al.*, 2018).

Research on polymerases and their interplay with modified substrates strongly indicate that both the position at which the modification is introduced, and the type of modification play crucial roles in the acceptance of a modified nucleotide by a polymerase. The incorporation efficiency also depends on the linker used to anchor the modification to the nucleobase, especially if sterically demanding groups are attached – typically, bulky modifications are accepted when being attached via a long flexible linker.

Natural polymerases exhibit some plasticity in substrate recognition, although obviously they have not been evolved by nature to tolerate unnatural nucleotides, thus it is intriguing how they are able to do so and what structural prerequisites might guide further engineering of improved enzyme variants. The studies unanimously report that DNA polymerases from family B are superior catalysts for unnatural substrates than members of the family A (Jäger *et al.*, 2005; Kuwahara *et al.*, 2006; Lapa *et al.*, 2016). As for the synthesis of modified RNA, it almost exclusively relies on the use of T7 RNA polymerase to avoid the need for complex transcription factors (Milisavljevič *et al.*, 2018).

1.2.1 Incorporation of modified substrates by DNA polymerases

The acceptance of base-modified nucleotides by DNA polymerases had been hardly predictable until first structural data of KlenTaq polymerase in complex with C5-modified pyrimidines came out in 2010. As compared to the structures with natural nucleotides, interaction with dUTP analog bearing nitroxide modification attached via rigid acetylene linker differs in the position of amino acid side chain Arg660, which is suggested to stabilize the closed and active conformations through hydrogen bonding with the phosphate backbone of the primer 3′ terminus. The Arg660 is substantially displaced as a result of steric hindrance of the bulky modification, which in turn may account for nearly 2500-fold decrease in incorporation efficiency. In contrast, modified nucleotide containing dendron anchored via the propargylamide linker, that is able to form hydrogen bonds with enzyme, was better substrate for the polymerase probably because of stabilization of the closed complex poised for catalysis. Despite the differences in Arg660 orientation, it seems that enzyme follows similar mechanisms to promote catalysis of polymerization of both natural and unnatural nucleotides, i.e. the formation of a stable clamp between the finger domain of the polymerase and the primer/template duplex is required for successful catalysis and may be achieved either via natural molecular contacts or via interactions with nucleotide modifications. Amino acid sequence alignment of several family A DNA polymerases revealed that the abovementioned arginine is conserved in bacteria, thus, it is likely that observed mechanism of enzyme-substrate complex stabilization applies to other DNA polymerases in this sequence family (Obeid *et al.*, 2010).

Researchers continued their exploration of KlenTaq structures with pyrimidine and purine analogs bearing longer modifications, such as (hydroxydecanoyl)-aminopentynyl. The crystal structures revealed that modifications extend outside the protein, taking different orientations: C5-position of pyrimidine orients the modification through the cavity formed by Arg587 and amino acid residues from the O-helix, while purine modification at C7-position was oriented above the 5′-triphosphate group through the cleft formed by residues from the palm domain, O-helix and Arg587 (Fig. 1.6). The possibility of a modification to extend to the outside of the protein through the described cavities enables enzymatic incorporation of even very bulky groups into DNA, if they are attached via a sufficiently long linker. All in all, these studies revealed that DNA polymerases interact with the modifications and stabilize unnatural conformations that may improve substrate properties. Positively charged amino acids, such as arginine and lysine, located near the active site undergo hydrogen bonding with functional groups of the modifications – this should be considered in the design of modified nucleotides to be efficient substrates (Bergen *et al*., 2012; Hottin & Marx, 2016; Hottin *et al*., 2017).

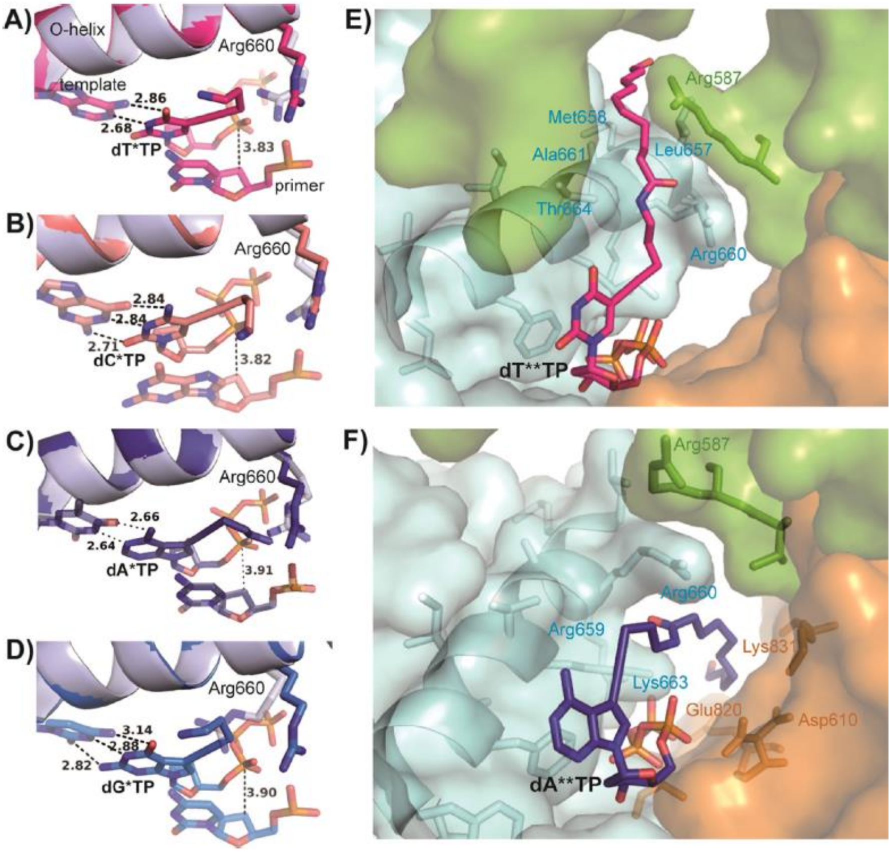


Figure 1.6. Structures of KlenTaq DNA polymerase bound to base-modified dNTPs. **(A-D)** – close-up views depicting KlenTaq DNA polymerase in complex with dT*TP (A), dC*TP (B), dA*TP (C) and dG*TP (D). **(E-F)** – close-up view of long modifications of dTTP (E) and dATP (F) pointing outside the protein. The finger, thumb and palm domains are indicated in pale blue, green and orange, respectively (by Hottin & Marx, 2016).

Kropp and co-workers further analyzed the elongation process on a structural level to understand how KlenTaq is able to execute postincorporation elongation from the modified nucleotide. Surprisingly, modifications adopted several distinct conformations, depending on their positioning in the primer – KlenTaq conformations were modulated by the modification and, in turn, the protein environment modulated the conformation of the modified nucleotide, all without compromising the enzyme′s activity. This indicates the remarkable plasticity of the system that may play role in the substrate properties of KlenTaq polymerase (Kropp *et al.*, 2018).

A selection of successful examples of the utilization of family A DNA polymerases for the incorporation of various base-modified nucleotides is provided in Table 1.1. For detailed descriptions of the structures of modified nucleotides, testing conditions and obtained incorporation efficiencies, please refer to the original publications.

Table 1.1. Family A DNA polymerases capable of base-modified nucleotide incorporation. The structure of a single representative – KlenTaq – is shown as a cartoon model (PDB code 4BWJ).

Family A	Polymerase	dN*TP	References
	KlenTaq	(Hydroxydecanoyl)-aminopentynyl-dATP/dUTP	Hottin *et al.*, 2017
		ON-dUTP	Baccaro *et al.*, 2012 Verga *et al.*, 2015
		Nitroxide-dUTP	Hollenstein, 2012
		Dendron-dUTP	
		HRP-dUTP	Welter *et al.*, 2016
	Taq	Biotin-dUTP	Anderson *et al.*, 2005
		AMCA-dUTP	
		Rhodamine-dUTP	
		Fluorescein-dUTP	
		Propynyl-dUTP	Kuwahara *et al.*, 2003
		Methyl-dCTP	
	Thermo Sequenase	Propynyl-dUTP	
		Methyl-dCTP	
		FAM-rhodamine-ddNTP	Kumar *et al.*, 2007
	DyNAzyme™	Ferrocene-dATP/dUTP	Hocek *et al.*, 2008
		Nitrophenyl-dATP/dUTP/dCTP	
	Klenow Fr. (exo-)	ON-dUTP	Verga *et al.*, 2015
	Klenow Fr.	Urea-dUTP	Hollenstein, 2012
		L-proline-dUTP	

	Sulfamide-dUTP	
	Ferrocene-dATP/dUTP	Hocek *et al.*, 2008
	Nitrophenyl-dATP/dUTP/dCTP	
	Biotin-dUTP	Anderson *et al.*, 2005
	AMCA-dUTP	

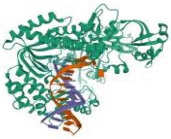
Interactions of unnatural substrates with the family B polymerases have been studied in less detail. The binary structure of KOD DNA polymerase revealed that the primer-template duplex adopts a B-form DNA conformation, with most 2′-deoxyribose moieties showing the ideal folding for B-DNA. In contrast, DNA observed near the insertion site of KlenTaq as well as other family A polymerases is in A-form conformation. B-form DNA is more elongated with a wide-opening major groove as compared to A-form DNA, thus DNA duplex conformation in family B DNA polymerases might favor the acceptance of modified substrates.

Comparison of the protein-DNA contacts between A and B family polymerases demonstrated some differences related to nucleobase contacts: six DNA nucleobases interact with five amino acid side chains in KlenTaq polymerase while only five nucleobases interact with three amino acid residues in KOD DNA polymerase. Further differences were found in the thumb domain – the tip of the thumb domain interacts with the primer strand in both types of polymerases, however the contact area in KOD polymerase structure is positioned above the minor groove whereas the corresponding area in KlenTaq extends into the major groove. As a result, nucleobase modifications located in the major groove clash with the tip of the domain when processed by KlenTaq. These observations might explain the better efficiency of family B polymerases in incorporating modified nucleotides (Hottin & Marx, 2016).

Wynne and colleagues sought to elucidate the structural details of mutant *Pfu* E10 polymerase, which was derived from an exonuclease-deficient *Pfu* variant using *in vitro* evolution techniques, in a complex with Cy5-modified dCTP (Wynne *et al*., 2013). *Pfu* E10 exhibits unique activity for high-density incorporation of cyanine-labeled dCTP (Ramsay *et al*., 2010). Although only apo form and its binary complex with DNA were obtained in this study, the modeling of the ternary complex suggested that bulky cyanine residue might be accommodated in the active center without any significant rearrangements of the enzyme amino acids, with the dye moiety located in the major groove of the duplex. Four mutations in *Pfu* E10 (E399D, N400D, R407I, Y546H) were attributed to those conferring the ability to incorporate cyanine-labeled substrates. Interestingly, particular variants of other polymerases from families A and B with an improved ability to accept unnatural substrates have mutations at equivalent positions.

A variety of family B DNA polymerases experimentally tested for base-modified nucleotide incorporation ability are listed in Table 1.2. It is worth noting that substantial part of adopted enzymes are exonuclease-deficient variants. Strong 3′-5′ exonuclease activity interferes with the ability to incorporate and extend nucleotide analogs presumably because proofreading ability results in the removal of unnatural nucleotides (Anderson *et al.*, 2005).

Table 1.2. Family B DNA polymerases capable of base-modified dNTP incorporation. The structure of a single representative – KOD – is shown as a cartoon model (PDB code 5OMF).

Family B	Polymerase	dN*TP	References
	KOD	$dU^{AA}TP^{(a)}$	Liu *et al.*, 2015
		Vinyl-dUTP/dCTP/dATP	Mačková *et al.*, 2014
		Antibody-dCTP	Balintová *et al.*, 2018
	Therminator™	$dU^{AA}TP^{(a)}$	Liu *et al.*, 2015
		ON-dUTP	Baccaro *et al.*, 2012
	Vent™ (exo-)	$dU^{AA}TP^{(a)}$	Liu *et al.*, 2015
		Thiophene-dUTP	Le *et al.*, 2017
		7-amino-2,5-dioxa-heptyl-dUTP/dCTP	Kuwahara *et al.*, 2003
		Urea-dUTP	Hollenstein, 2012
		L-proline-dUTP	
		Sulfamide-dUTP	
		Vinyl-dUTP/dCTP/dATP	Mačková *et al.*, 2014
		Biotin-dUTP	Anderson *et al.*, 2005
		AMCA-dUTP	
		Rhodamine-dUTP	
		Fluorescein-dUTP	
	DeepVent™ (exo-)	Iodo-dCTP	Whitfield *et al.*, 2018
		Octadiynyl-dCTP	
		Bromo-dUTP	
		7-deaza-7-iodo-dATP	
		7-amino-2,5-dioxa-heptyl-dUTP/dCTP	Kuwahara *et al.*, 2003

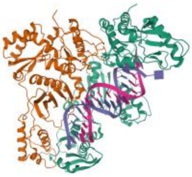
	Biotin-dUTP	Anderson *et al.*, 2005
	AMCA-dUTP	
	Rhodamine-dUTP	
	Fluorescein-dUTP	
Tgo-Pol-Z3 (exo-)	Iodo-dCTP	Whitfield *et al.*, 2018
	Octadiynyl-dCTP	
	Bromo-dUTP	
	7-deaza-7-iodo-dATP	
9°N	PEG-dUTP	Baccaro *et al.*, 2010
	Dendron-dUTP	
Pwo	PEG-dUTP	Baccaro *et al.*, 2010
	Dendron-dUTP	
	7-amino-2,5-dioxa-heptyl-dUTP/dCTP	Kuwahara *et al.*, 2003
	Urea-dUTP	Hollenstein, 2012
	L-proline-dUTP	
	Sulfamide-dUTP	
Pwo	Vinyl-dUTP/dCTP/dATP	Mačková *et al.*, 2014
	Phenylalanine-dATP/dUTP	Hocek *et al.*, 2008
Pfu	7-amino-2,5-dioxa-heptyl-dUTP/dCTP	Kuwahara *et al.*, 2003
	Cy3-dCTP	Ramsay *et al.*, 2010
	Cy5-dCTP	

[a] $\mathrm{dU^{AA}TP}$ stands for 5-amino-dUTP bearing amino acid-like functional groups.

Terminal deoxynucleotidyl transferase (TdT) – a family X DNA polymerase – was reported to be substrate promiscuous, especially in the presence of cacodylate buffer and cobalt ions (Hatahet *et al.*, 1993), although early reports on enzyme compatibility with base-modified substrates focused on nucleotide analogs with small modifications (Motea & Berdis, 2010). Sørensen *et al.* proposed to employ TdT for template-independent direct ligation of polymers, proteins and other large biomolecules to the 3′ terminus of any native ON (Sørensen *et al.*, 2013). The crystal structure of TdT reveals how the enzyme is able to accept nucleotides with bulky modifications on the nucleobase. The nucleobase is facing towards a wide, open crevice near the outer sphere of the enzyme. This might allow binding of the macromolecule-conjugated nucleotide at the active site without significant steric interference. Moreover, the substrate DNA strand is oriented away from the open crevice in a way that would also not cause significant steric hindrance.

Reports on compatibility of other DNA polymerases, such as reverse transcriptases, with unnatural substrates are scarce, probably due to the limited repertoire of applications which would require biochemical properties not exhibited by enzymes of families A, B and X. Nevertheless, reverse transcriptases which were included in screening studies demonstrated the ability to use base-modified analogs. The examples of modified nucleotides that were shown to serve as substrates for TdT and RTs are listed in Table 1.3.

Table 1.3. DNA polymerases from families X and RT capable of base-modified dNTP incorporation. The structures of representatives are shown as cartoon models. PDB codes are 4I29 (mouse TdT) and 6HAK (HIV-1 RT).

Family		**Polymerase**	**dN*TP**	**References**
X		TdT	Peptide-dCTP	Sørensen *et al.*, 2013
			PEG-dUTP	
			Dendrimer-dUTP	
			Streptavidin-ddUTP	
RT		M-MLV RT	Biotin-dUTP	Anderson *et al.*, 2005
			AMCA-dUTP	
			Rhodamine-dUTP	
			Fluorescein-dUTP	
		AMV RT	Biotin-dUTP	
			AMCA-dUTP	
			Fluorescein-dUTP	

1.2.2 Incorporation of modified substrates by RNA polymerases

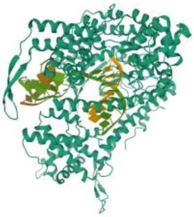
Both chemical and enzymatic methods for synthesis of modified RNA are less well established despite the extensive efforts in the use of modified RNA probes for imaging, chemical biology and therapeutic applications (Anhäuser & Rentmeister, 2017). The enzymatic synthesis can be based either on the incorporation of modified NTPs into RNA or on enzymatic posttranscriptional modifications of RNA, e.g. alkylation by methyltransferases. Alternatively, posttranscriptional chemical modifications might be introduced via bioorthogonal chemistry (Milisavljevič *et al.*, 2018).

T7 RNA polymerase (T7 RNAP) is one of the simplest enzymes known to synthesize RNA – it produces RNA transcripts from a dsDNA template without the use of additional transcription factors, which makes it ideal for *in vitro* applications (Vaught *et al.*, 2004). T7 RNAP requires specific promoter and the presence of guanosines in the +1, +2 and/or +3 positions to ensure efficient transcription initiation (Rong *et al.*, 1998; Kennedy *et al.*, 2007).

Modified nucleotides bearing small base modifications were reported to be almost as good substrates for T7 RNAP as natural nucleotides, while bulky modifications resulted in less efficient transcription. The known examples of T7 RNAP incorporation of base-modified substrates are summarized in Table 1.4. Because of the inherent dependency of transcription initiation on the presence of GTPs, processing of base-modified GTPs is more difficult – it seems that T7 RNAP does not tolerate unnatural bulky modifications at the +1 position (Milisavljevič *et al.*, 2018).

T7 RNAP is specialized in synthesizing RNA rather than DNA strands. Previous studies indicated that wild type enzyme efficiently discriminates between NTPs and dNTPs. Structural analysis suggested that substrate selection occurs in the T7 RNAP preinsertion site through Mg^{2+}-mediated interaction of the Tyr639 hydroxyl with the 2′OH group of substrate ribose. To allow the nucleotide incorporation, Tyr639 has to move out of the active site during the insertion process. When dNTP occurs in the preinsertion site of the wild type RNAP, Tyr639 associates with dNTP under the mediation of localized water molecules and its side ring stacks strongly to the end base pair of the RNA-DNA hybrid at the 3′ end of the nascent RNA strand resulting in the blockage of the active site and rejection of dNTP. The role of Tyr639 residue in substrate selection process was further illustrated upon characterization of T7 RNAP Y639F mutant, which exhibits 20-fold increase in dNTP incorporation relative to wild type enzyme but maintains the wild type level of activity in the NTP incorporation (Sousa & Padilla, 1995; Temiakov *et al.*, 2004; Duan *et al.*, 2014).

Table 1.4. Base-modified NTPs reported to be substrates of T7 RNA polymerase. The structure of the enzyme is shown as a cartoon model (PDB code 1H38).

T7 RNAP	N^*TP	References
	Biotin-UTP	Langer *et al.*, 1981
	Vinyl-UTP	George *et al.*, 2017
	Iodo-UTP	Walunj *et al.*, 2018
	Phenyl-UTP	Vaught *et al.*, 2004
	4-pyridyl-UTP	
	2-pyridyl-UTP	
	Indolyl-UTP	
	Isobutyl-UTP	
	Imidazole-UTP	
	3-aminopropyl-UTP	Vaish *et al.*, 2000
	Diazirine-UTP	Smith *et al.*, 2014

Naphthalimide-UTP	Tanpure *et al.*, 2014
Furan-UTP	Srivatsan *et al.*, 2007
Benzo[*b*]thiophene-UTP	Pawar *et al.*, 2011
Thiophene-UTP	Srivatsan *et al.*, 2009
Benzofuran-UTP	Tanpure *et al.*, 2011
Ferrocene-UTP	Di Giusto *et al.*, 2004
Anthraquinone-UTP	
Trans-cyclooctene-CTP	Asare-Okai *et al.*, 2014
2-thienyl-7-deaza-ATP	Perlíková *et al.*, 2016
7-ethynyl-8-aza-7-deaza-ATP	Zheng *et al.*, 2016
Methyl-ATP/GTP	Milisavljevič *et al.*, 2018
Ethynyl-ATP/UTP/CTP	
Phenyl-ATP/UTP/CTP	
Benzofuryl-ATP/UTP/CTP	
Dibenzofuryl-UTP/CTP	

Certain transcriptomic applications rely on the analysis of poly(A) tail related events and benefit from the opportunity to label 3′ termini of RNA with ATP analogs (Curanovic *et al.*, 2013). Zheng and co-workers have demonstrated that 7-ethynyl-8-aza-7-deaza-ATP is substrate not only for T7 RNAP, but also for *Escherichia coli* poly(A) polymerase (Zheng *et al.*, 2016). Such observations further expand the toolbox of nucleotide analogs and enzymes useful for RNA labeling.

1.2.3 Biocompatibility of modified nucleic acids

An essential requirement for the use of modified nucleic acids in molecular biology is that the modifications should be benign, with the modified nucleic acids being a functional mimic of their natural counterparts (Sanzone *et al.*, 2012). Relatively few studies have focused on replication or transcription through artificial backbones. It was reported that minor phosphodiester modifications are accepted by polymerases (Ciafrè *et al.*, 1995), an amide variant was imperfectly bypassed in primer extension experiments (Kuwahara *et al.*, 2009), and certain triazole-based backbones are functional *in vitro* and even *in vivo* in bacterial and mammalian cells (El-Sagheer *et al.*, 2011; Birts *et al.*, 2014).

Shivalingam and colleagues aimed to understand molecular requirements for high-fidelity replication of artificial DNA backbones (Shivalingam *et al.*, 2017). They have synthesized ONs containing several structurally and electronically varied artificial linkages based on triazole, amide and phosphorothioate modifications. The results of linear primer extension and

PCR amplification showed that even minor phosphorothioate modifications can impair the copying process while some radical triazole and amide backbones performed surprisingly well, indicating that the phosphate group itself is not essential. Reading through the backbone linkage was identified as a rate limiting step in copying process: the speed of replication correlated well with the steric demands of respective artificial backbones. Phusion™ (exo+) DNA polymerase generally replicated artificial linkages more efficiently than *Taq* (exo-) polymerase at shorter extension times. As these times were lengthened, product yields for Phusion did not increase as significantly as for *Taq*.

The authors have assessed the fidelity of replication by NGS. Strikingly, despite being far more similar to natural phosphodiester backbone than other studied modifications, phosphorothioate linkages showed significant insertions, the length and position of which depended upon the polymerase used for copying. Strong interaction between the sulfur atom and the polymerase may have inhibited polymerase passage through the template thereby promoting multiple dNTP additions.

Multibase deletions were more prominent for polymerases exhibiting proofreading activity. It was suggested that exo+ polymerase stalls at the modification site and passes the primer terminus to the 3′-5′ exonuclease site which arbitrarily digests the extended primer. This extension and digestion process may continue until either the modification is passed, or the polymerase loops the modified backbone out of the template to enable its unimpeded extension. For exo- polymerases, the looping mechanism is only accessed once, thus reducing the possibility of multibase deletions. This phenomenon appears to be also linked to sugar distortions – lower-level multibase deletions were observed for amide backbone which does not significantly perturb sugar placement or conformation.

The point deletions around the artificial linkage correlated well with the hybridization of the atom immediately adjacent to the 5′-3′-side nucleosides (sp^2 *vs* sp^3), and poorly – with internucleoside bond separation or the backbone functional group. This observation is consistent with the known phenomenon that polymerases bend the backbone of the template by ~90° immediately 5′ to the site of dNTP addition (Arias-Gonzalez, 2017), indicating the importance of backbone flexibility.

Overall, this study elucidated general prerequisites for biocompatibility of artificial backbones (Fig. 1.7) and demonstrated that certain triazole linkers are good DNA backbone analogs, which is encouraging for the use CuAAC “click” ligation as means to assemble biocompatible modified DNA.

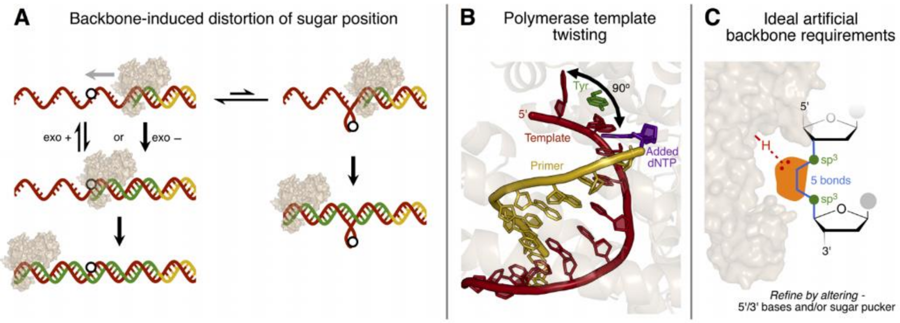

A
Backbone-induced distortion of sugar position
exo +
or
exo –
B
Polymerase template twisting
Tyr
90°
5'
Template
Primer
Added dNTP
C
Ideal artificial backbone requirements
5'
H
sp3
5 bonds
sp3
O
3'
Refine by altering - 5'/3' bases and/or sugar pucker

Figure 1.7. Suggested features that the ideal artificial backbone must avoid or fulfill. **(A)** – the mechanism by which backbone-induced distortion of the sugar position might generate multibase deletions around the linkage. **(B)** – the 90° twist in template geometry induced by polymerases to enable dNTP addition. **(C)** – the requirements of an artificial backbone to enable the mechanism depicted in part (B) and thereby allow accurate recognition of the adjacent 5′/3′ bases (by Shivalingam *et al.*, 2017).

Many of the observed effects were highly dependent on the polymerase, suggesting that it may be possible to artificially evolve polymerases to perfectly accommodate modified DNA linkages.

1.3 High-throughput techniques for nucleic acid analysis

We are witnessing the rapid development of next generation sequencing (NGS) technologies for almost two decades. NGS allows not only faster high-throughput analysis of disease-related genes at lower costs than traditional approaches (D'Argenio *et al.*, 2015), but also the sequencing of panels of genes up to complete exomes or genomes (Miller *et al.*, 2017). In this way, it is possible to increase the diagnostic sensitivity, to discover novel disease-related genes and obtain data on other genes potentially acting as disease-phenotype modifiers (Weber *et al.*, 2016). Due to higher sensitivity and flexibility, NGS techniques are also useful for prenatal diagnostics (Maxwell *et al.*, 2016) and other applications, such as sequencing of circulating free DNA (Huang *et al.*, 2017).

In transcriptomics, NGS-based approaches have outcompeted the use of microarrays by allowing the analysis of virtually all RNA molecules, known and unknown, present in a sample, at a lower cost (Sandhu *et al.,* 2018). Moreover, alternative splicing isoforms, fusion genes, long non-coding transcripts and small RNAs can be sequenced and analyzed both structurally and quantitatively in a hypothesis-free manner (Precone *et al*., 2015; Chen *et al*., 2018; Ilott & Ponting, 2013; Motameny *et al*., 2010). In addition, recent technological advances are showing the tremendous potential of NGS for single cell analysis (Hwang *et al*., 2018).

NGS has prompted the high-throughput studies of the epigenome and of the microbiome. By using the specific library preparation protocols, it is possible to analyze the methylation status of DNA at a genome-wide level or by focusing on a custom set of genomic regions of interest (Widschwendter *et al*., 2017; Pu *et al*., 2017). Moreover, chromatin immunoprecipitation sequencing approaches have shown their efficacy in the studies of the regulatory networks of gene expression at the genome-wide level, by allowing the identification of the targets of specific transcription factors (Pavesi, 2017). NGS-based techniques gave a significant boost to metagenomics for the study of microbial relationships with human physiology and pathology, and for the identification of specific microbial signatures related to a disease of interest, by eliminating the need of microbial cultivation (Malla *et al*., 2019). The most impressive result of the first studies that sequenced all DNA molecules within a metagenomic sample was that up to ten times more organisms were encountered than seen previously, giving an idea of the complexity and constitution of entire ecosystems (Venter *et al*., 2004).

As the costs of NGS continue to decrease, it is conceivable to hypothesize that these and many other applications will become even more common and will be eventually implemented into clinical practice (D'Argenio, 2018).

All currently available sequencing platforms require some level of nucleic acid pre-processing into a library suitable for sequencing. Generally, these steps include fragmentation of DNA or RNA into an appropriate platform-specific size range, followed by end polishing to generate termini suitable for ligation. Specific adapters are then attached to these fragments. A functional library requires specific adapter sequences to be added to the 3′ and 5′ termini (Buermans & den Dunnen, 2014) and then to be amplified by PCR, if necessary.

Commonly used adapter addition methods are based on enzymatic ligation or introduction via PCR (Menzel *et al*., 2014). As PCR may introduce artifacts through stochasticity, template switches and polymerase errors (Kebschull & Zador, 2015), ligation-based PCR-free library preparation methods are preferred if the amount of starting material permits to obtain sufficient library yield (Huptas *et al*., 2016). Nevertheless, ligation also distorts the original library composition because of sequence- or secondary structure-dependent biases of DNA and RNA ligases (Zhuang *et al*., 2012; Seguin-Orlando *et al*., 2013). Moreover, ligation efficiency obtainable by commercially available kits varies between 3-20%, which further reduces library complexity (Aigrain *et al.*, 2016).

Next chapters will focus on demonstrated DNA and RNA library preparation techniques that benefit from the use of chemically modified nucleic acids and chemical ligation.

1.3.1 DNA sequencing

Efficient adapter tagging of ssDNA is of interest for many DNA-seq applications: methods originally developed for the genomic analysis of highly degraded ancient DNA (Gansauge & Meyer, 2013) have later been adopted for other fragmented sample types, such as cell-free DNA and DNA purified from formalin fixed paraffin embedded sections. Among the advantages of such approaches is the correspondence of sequencing reads to the natural 5′ and 3′ ends so that data mapped to the reference genome reveals the exact genomic locations of the input fragments, which is an important feature for researchers studying biological fragmentation patterns (Troll *et al.*, 2019).

Miura and co-workers developed an alternative chemoenzymatic approach termed TdT-assisted, CuAAC-mediated ssDNA ligation, or TCS ligation (Miura *et al.*, 2018). Here, TdT is used to incorporate a single 3′-azide-modified dideoxynucleotide onto the 3′ end of target ssDNA, followed by CuAAC-mediated “click” ligation of the azide-incorporated 3′ end to a 5′-ethynylated synthetic adapter. After second strand synthesis, the second synthetic adapter is added by T4 DNA ligase (Fig. 1.8, A). The authors were able to optimize the efficiency of chemical ligation up to 20-30%. An important task was to find conditions for the efficient second strand synthesis through the triazole linkage: out of 16 tested DNA polymerases 14 failed to produce extension products in PEX assays, and two (Klenow fragment and Klenow fragment exo-) exhibited adequate read-through activity. Proof-of-principle library preparation experiments were conducted on a mixture of synthetic ONs. Libraries were successfully obtained, although 12-step protocol yielded conversion efficiency of only 0.6%.

Upon sequencing, several other undesirable effects were observed. First, more than a half of the reads (57.8%) were shorter than the expected 102 nt, which was attributed to the degradation of DNA during CuAAC ligation (Fig. 1.8, B). In addition, the nucleotide composition was somewhat distorted from the average at the beginning of sequencing read (Fig. 1.8, C). Because no similar deviations were visible in the control library, the 3′-proximal region of the target DNA likely influenced the efficiency of the TCS ligation-based workflow. The authors further demonstrated the applicability of their approach for MNase-seq, however the need for significant optimization of conversion efficiency was noted.

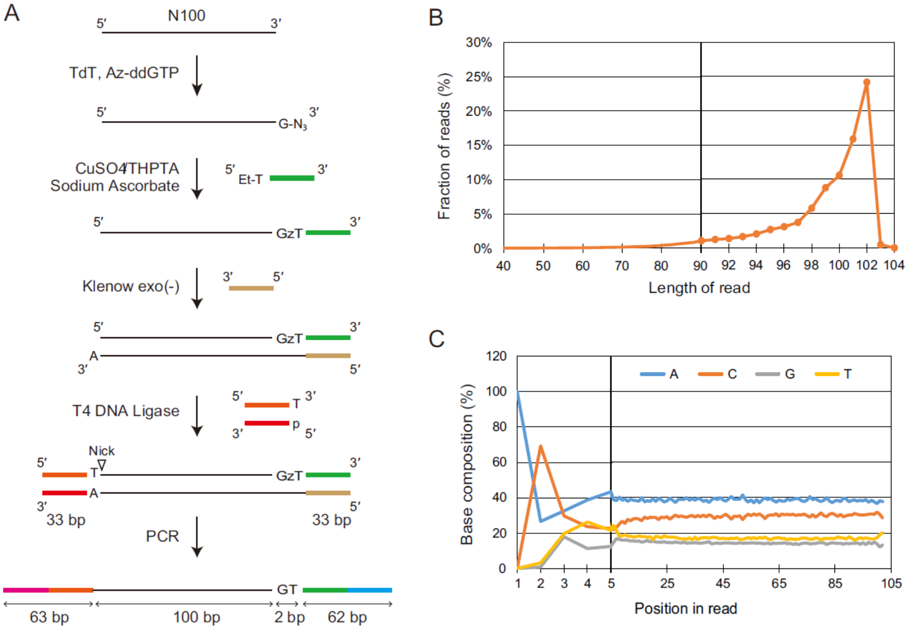


Figure 1.8. Library preparation from ssDNA. **(A)** – schematic overview of the library preparation workflow, using 100-mer ONs as input. **(B)** – read length distribution after trimming of adapter sequences. **(C)** – mean base composition of reads at each position. Sequencing was performed starting from the chemical ligation site (by Miura *et al.*, 2018).

Chemical ligation appeared to be instrumental for the development of epigenome profiling methods. Staševskij and colleagues introduced a method for high-resolution profiling of unmodified CG sites, termed tethered oligonucleotide-primed sequencing, or TOP-seq (Staševskij *et al*., 2017). This technique involves selective tagging of unmodified genomic CG sites with an azide group using an engineered variant of the SssI methyltransferase and a synthetic analog of the SAM cofactor. Next, alkyne-bearing DNA oligonucleotide is chemically tethered to azide-tagged loci employing "click" chemistry. This enables the tethered ON-primed synthesis of the complementary DNA strand in the proximity to the target site, obtaining nested DNA strands that sequentially include the CG site and its adjacent genomic region (Fig. 1.9).

CuAAC ligation in TOP-seq protocol was highly efficient as assessed by HPLC-MS analysis of reaction products obtained in a model system. The synthesis of a TO-primed complementary strand was executed using *Pfu* DNA polymerase. Although the reaction yielded fragments of a correct structure, which were able to participate in subsequent indexing PCR, the efficiency of read-through was not reported. TOP-seq was applied for the analyses of bacterial and human genomes and showed better agreement with published bisulfite sequencing datasets as compared to widely used MBD-seq (Serre *et al.*, 2010) and MRE-seq (Maunakea *et al.*, 2010) methods. TOP-seq offered an appealing combination of single CG resolution, genome-wide coverage and affordable cost, expanding the toolbox for high-throughput epigenome profiling.

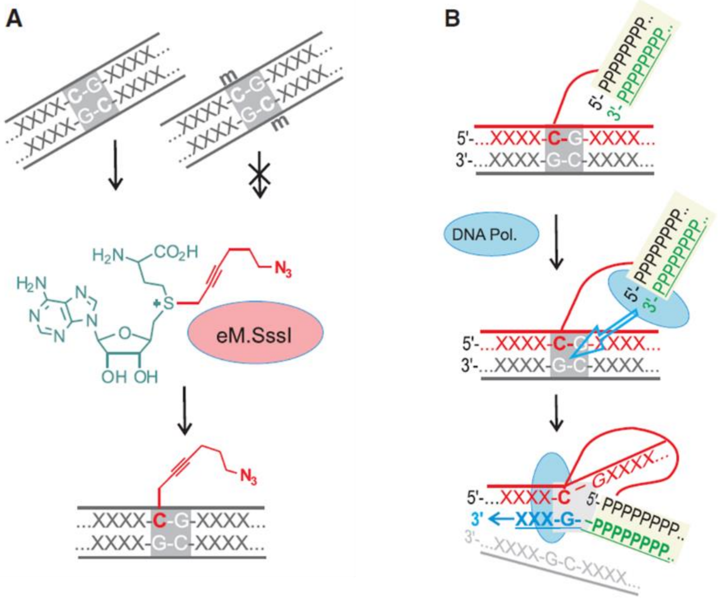


Figure 1.9. Overview of the TOP-seq approach. **(A)** – selective tagging of unmethylated CG sited with an azide group using engineered SssI methyltransferase (eM.SssI) and a synthetic SAM analog. **(B)** – tethered oligonucleotide-primed DNA polymerase activity at an internal covalently tagged CG site (by Staševskij *et al.*, 2017).

1.3.2 RNA sequencing

Routh and co-workers explored the possibility to exploit “click” chemistry to simplify RNA-seq library preparation workflow (Routh *et al.*, 2015). In the approach termed ClickSeq, the authors performed randomly primed reverse transcription reactions supplemented with azido-2′,3′-dideoxynucleotides that stochastically terminated cDNA synthesis and generated 3′-azido blocked cDNA fragments in a process similar to Sanger sequencing. Purified fragments were then "click"-ligated via CuAAC to DNA ONs modified with 5′-alkyne group. This resulted in ssDNA molecules containing an unnatural triazole-linked DNA backbone that was found to be compatible with PCR amplification.

The efficiency of chemical ligation was estimated to be ~10%, and subsequent reading through an unnatural linkage by *Taq* DNA polymerase was found to be even less efficient (<4%), but nevertheless the libraries of randomly distributed fragments covering viral genomes were obtained and sequenced. The authors reported lower rates of artifactual recombination in ClickSeq data as compared to other commercially available kit. This observation was attributed to the removal of the fragmentation step and selectivity of chemical ligation which prevented the formation of RNA fragments able to ligate to one another; and the fact that 3′-azido blocked cDNAs cannot form a priming substrate for artefactual template switching. Low chimera rates allowed confident detection of natural recombination events, which is a valuable feature for diverse areas of research, including mRNA splicing, detection of chromosomal rearrangements and others.

Later, the same group published an altered version of ClickSeq, called poly(A)-ClickSeq or PAC-seq (Routh *et al.*, 2017). This approach uses poly(T) reverse transcription primer instead of random ONs, which leads to the construction of RNA-seq libraries enriched for 3′UTR/poly(A) junctions (Fig. 1.10). The authors envisioned numerous advantages of employing 3′ end sequencing for characterizing quantitative changes in the transcriptome (Elrod *et al.,* 2019):

- Library complexity is limited to one fragment per transcript. This saves on the amount of sequencing that must be performed as compared to standard RNA-seq covering the whole transcript length.
- As transcripts have only one poly(A) tail, this negates the need for computation normalization of read counts assigned to mRNA as a function of their length.
- Short transcripts that would otherwise receive very low sequence coverage in standard RNA-seq can be accurately quantified in an equivalent manner to longer transcripts.
- Only mature mRNA transcripts that contain long poly(A) tails are captured, thus allowing accurate representation of the translating mRNAs.

A notable limitation of 3′ end sequencing is possible priming from A-rich sequences within mRNAs. This may result in absolute read counts being elevated for particular transcripts. However, the authors expect that the frequency of internal priming from A-rich regions should correlate with transcript abundance and be conserved among multiple replicates. In this case, internal priming may not excessively perturb differential gene expression analysis.

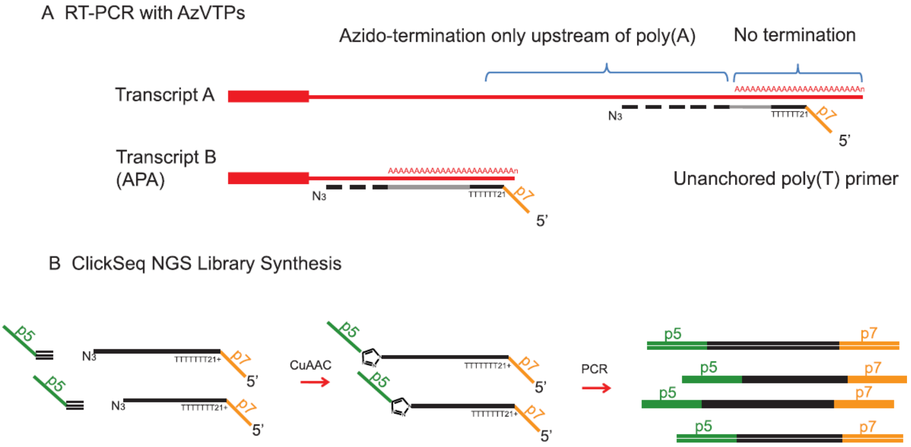


Figure 1.10. An overview of poly(A)-ClickSeq technique. **(A)** – RT-PCR is initiated from a non-anchored poly(T) primer containing a portion of the Illumina P7 adapter. The reaction is performed in the presence of azido-modified ddATP, ddGTP and ddCTP. **(B)** – azido-blocked cDNA fragments are “click”-ligated to 5′-hexynyl-functionalized DNA ONs containing the P5 Illumina adapter (by Routh *et al*., 2017).

To illustrate the utility of PAC-seq, the authors depleted a component of the *Drosophila* Integrator complex in DL1 cells using RNA interference technique and compared the changes of gene expression relative to control using both standard RNA-seq and PAC-seq. In addition to providing information on the position of poly(A) tail, PAC-seq revealed global changes in mRNA transcript abundance which closely matched those observed in RNA-seq data. From technical perspective, PAC-seq protocol proved to be feasible with total RNA inputs down to 125 ng and generated ~50% of usable reads.

Recently, Mikutis and colleagues proposed an elegant approach for epitranscriptomic sequencing which relies on “click” chemistry (Mikutis *et al*., 2020). The group reported the development of click-degraders – small molecules that can be covalently attached to RNA via “click” reaction and can degrade them like ribonucleases. Click-degraders have become the basis of meCLICK-Seq (methylation CLICK-degradation sequencing) method useful for identification of RNA modification substrates with high resolution at intronic and intergenic regions. The technique hijacks RNA methyltransferase activity to introduce an alkyne, instead of methyl, group on RNA. Subsequent CuAAC reaction with the click-degrader leads to targeted RNA cleavage (Fig. 1.11).

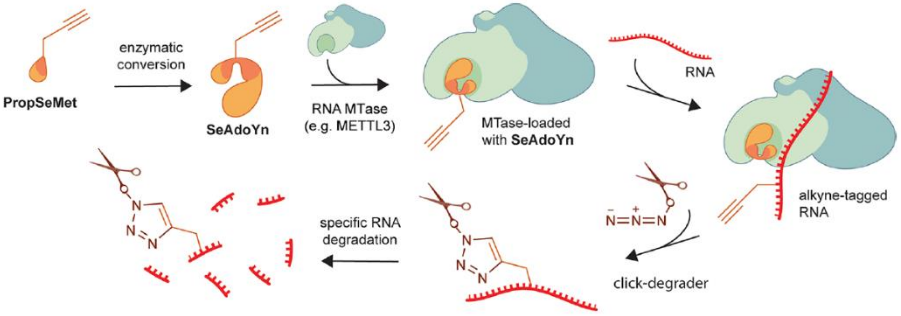


Figure 1.11. The proposed mechanism of action of meCLICK-Seq, a small molecule-based methylated RNA editing platform (by Mikutis *et al*., 2020).

Impressively, meCLICK-Seq successfully applied “click” chemistry directly on live cells, with a quantifiable output. The authors note that unlike antibody-based methods, meCLICK-Seq does not require large quantities of RNA, does not rely on the availability of an antibody against a particular modification and does not involve enzymatic or any other kind of *in vitro* RNA processing prior to library prep. Moreover, the technique depends strictly on the catalytic activity of RNA methyltransferases, such that it can determine their transcript and locus specificity.

The applicability of meCLICK-Seq was illustrated by identifying transcript substrates of N^6-methyladenosine (m^{6}A) writers METTL3 and METTL16 in human MOLM-13 cells treated with methionine surrogate PropSeMet. Moreover, the authors demonstrated that m^{6}A is widespread in long non-coding RNAs as well as in intronic and intergenic regions.

These early examples open doors for the introduction of chemoenzymatic methods of nucleic acid processing into established and emerging high-throughput nucleic acid analysis techniques. The interplay between chemistry and molecular biology allows not only simplification of protocols, but also gaining previously inaccessible biological insights.

2. MATERIALS AND METHODS

All oligonucleotides used in this work were synthesized by Metabion GmbH requesting HPLC purification.

2.1 Synthesis of oligonucleotide-tethered dideoxynucleotides

All reaction components were added to the reaction mixture as solutions in nuclease-free water unless specified otherwise. Modified oligonucleotides used for coupling to dideoxynucleotides are listed in Table 2.1.

Table 2.1. Modified oligonucleotides used in this study

Name	Oligonucleotide sequence
ON1	5′-(AldU)-AGATCGGAAGAGCACACGTCTG-biotin-3′
ON2	5′-hexynyl-AGATCGGAAGAGCACACGTCTG-biotin-3′
ON3	5′-hexynyl-AGATCGGAAGAGCACACGTCTG-pho-3′
ON4	5′-hexynyl-AGATCGGAAGAGCACACGT*C*T*G-pho-3′
ON5	5′-hexynyl-NNNNNNNNAGATCGGAAGAGCACACGTCTG-biotin-3′
ON6	5′-hexynyl-NNNNNNNNAGATCGGAAGAGCACACGT*C*T*G-pho-3′
ON7	5′-(AldU)-NNNNNNNNAGATCGGAAGAGCGTCGTGTA-biotin-3′
ON8	5′-hexynyl-NNNNNNNNAGATCGGAAGAGCGTCGTGTA-biotin-3′
ON9	5′-hexynyl-NNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAG-pho-3′
ON10	5′-(AldU)-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACATGCCTAAA TCTCGTATGCCGTCTTCTGCTTG-biotin-3′

„AldU“ stands for 5-(octa-1,7-diynyl)-dUTP, „pho“ – phosphate, * - phosphorothioate bonds.

Oligonucleotides were conjugated to azido-modified dideoxynucleotides using “click” chemistry. Dideoxynucleotide precursors were either 5-(3-(2-azidoacetamido)prop-1-ynyl)-2′,3′-dideoxypyrimidine-5′-triphosphates or 7-deaza-7-(3-(2-azidoacetamido)prop-1-ynyl)-2′,3′-dideoxypurine-5′-triphosphates. ddC^{N3}TP, ddU^{N3}TP, ddG^{N3}TP or ddA^{N3}TP (3 eq.) solution was added to 5′-alkynyl modified oligonucleotide (200-210 nmol) solution in sodium phosphate buffer (1 mL, 100 mM, pH 7). A premixed solution of $CuSO_4$ (100 mM, 12 eq.) and THPTA (250 mM, 5 eq. to $CuSO_4$) was then added to the reaction mixture, followed by the addition of sodium ascorbate (1 M, 50 eq. to $CuSO_4$). Reaction mixture was stirred for 20 min at 42°C, quenched with 0.5 M EDTA-Na_2 solution (1 ml, pH 8). The products were purified by C18 reversed-phase chromatography using 100 mM TEAAc/ACN

(10-30%, depending on the nature of azido-modified dideoxynucleotide and ON modification) as eluent and desalted using water/ACN (0-100%) as eluent.

Oligonucleotide-tethered dideoxynucleotides (OTDDNs or ddN^{ON}TPs) were typically obtained with >20% yield and >95% purity. The purity of obtained products was evaluated by HPLC, and molecular mass was verified by LC-MS. The synthesis principle and the structures of studied ddN^{ON}TPs are shown in Fig. 2.1.

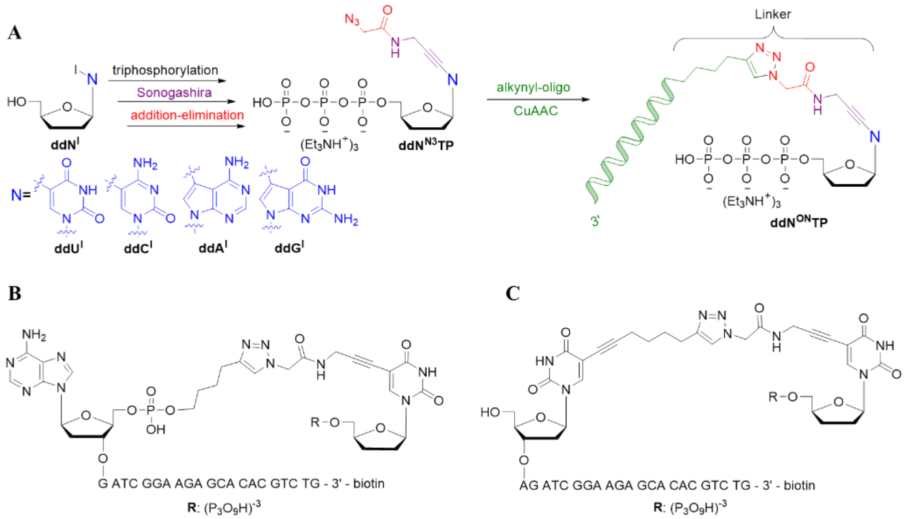


Figure 2.1. (A) – a brief depiction of ddN^{ON}TP synthesis principle. „N" stands for a nucleobase. **(B)** – the structure of ddN^{ON}TP which ON modification is attached using hexynyl on the 5′-terminal phosphate group as exemplified by ON2 oligonucleotide conjugate with azido-ddUTP (ddU^{ON2}TP). **(C)** – the structure of ddN^{ON}TP which ON modification is attached using alkyne moiety on the 5′-terminal nucleobase as exemplified by ON1 conjugate with azido-ddUTP (ddU^{ON1}TP).

2.2 Assays to test incorporation and read-through

A selection of various DNA polymerases, including representatives from family A, family B, family X and reverse transcriptase (RT) family, as well as RNA polymerases were tested for capability of incorporating OTDDNs. The experimental system for testing of DNA polymerases was based on the filling of 5′-protruding ends of the oligonucleotide duplexes. RNA polymerases were tested in *in vitro* transcription reaction using plasmid template containing appropriate promoter sequence or using 100 nt transcript for template-independent tailing.

Templates for incorporation testing. To prepare oligonucleotide duplexes with different protruding ends, oligonucleotides listed in Table 2.2

were annealed in 1× annealing buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 50 mM NaCl) to obtain 2 μM final solution. Primers were purchased with a fluorescent dye modification for subsequent detection of PEX products on a gel.

Table 2.2. Oligonucleotide duplexes used for OTDDN incorporation testing.

Name	Oligonucleotide duplex
Dup^{A}	5′-AAAAAAAAAATACGCCAAGGATGCCTACCCATGTCTGCA-3′ 3′-ATGCGGTTCCTACGGATGGGTACAGACGT-Cy5-5′
Dup^{T}	5′-TTTTTTTTTTTACGCCAAGGATGCCTACCCATGTCTGCA-3′ 3′-ATGCGGTTCCTACGGATGGGTACAGACGT-Cy5-5′
Dup^{G}	5′-GGGGGGGGGGTACGCCAAGGATGCCTACCCATGTCTGCA-3′ 3′-ATGCGGTTCCTACGGATGGGTACAGACGT-Cy5-5′
Dup^{C}	5′-CCCCCCCCCCTACGCCAAGGATGCCTACCCATGTCTGCA-3′ 3′-ATGCGGTTCCTACGGATGGGTACAGACGT-Cy5-5′
Dup^{N}	5′-GTCGCTCAACTCAGCTACAGTACGCCAAGGATGCCTACCCATGTCTGCA-3′ 3′-ATGCGGTTCCTACGGATGGGTACAGACGT-Cy5-5′

PEX reaction conditions for incorporation testing. Primer extension reactions were performed in commercial buffers and optimal or near-optimal temperatures for each tested polymerase. Control reactions with native dNTPs were conducted to ensure that polymerase of interest is capable to perform conventional primer extension at given conditions.

Tested polymerases and specific reaction conditions are listed in Table 2.3. In all cases, 2 pmol of oligonucleotide duplex (or single-stranded primer for TdT) and 20 pmol of OTDDN or corresponding native dNTP were used per reaction. Those polymerases which exhibited good OTDDN incorporation capability were later tested for competitive incorporation of OTDDN when modified terminators were mixed with native dNTPs in 1:1, 1:2 or 1:3 ratios, respectively. To test polymerases which do not exhibit 3′-5′ exonuclease activity, dideoxynucleotide conjugates with ON2 were used, whereas for testing of proofreading enzymes, conjugates with ON4 were used so that oligonucleotide modification would be resistant to polymerase-mediated degradation.

All polymerases and reaction buffers are manufactured by Thermo Fisher Scientific unless specified otherwise.

Table 2.3. Polymerases tested for OTDDN incorporation

Family	Polymerase, amount per reaction		Conditions
A	*Taq* DNA polymerase	2.5 U	95°C 1 min → 60°C 30 min
	Taq (exo-)	2.5 U	
	Tth DNA polymerase	2.5 U	
	Platinum™ II *Taq*	2.5 U	
	DyNAzyme™ II	2 U	
	DyNAmo™ IV	1.2 U	
	Thermo Sequenase	40 U	
	CycleSeq™	32 U	
	KlenTaq	2.5 U	
	Sequenase™ V2.0	13 U	37°C 30 min
	T7 DNA polymerase	10 U	
	Klenow fragment (exo-)	5 U	
	Bsm DNA polymerase	8 U	
	DNA polymerase I	10 U	
B	Platinum™ SuperFi™	2 U	95°C 1 min → 60°C 30 min
	Phusion (exo-)	2 U	
	Pfu (exo-)	2 U	
	Phusion U	2 U	
	T4 DNA polymerase	1 U	37°C 30 min
	Phi29 polymerase	10 U	
X	TdT	30 U	37°C 40 min
RT	Maxima™ RT	200 U	50°C 30 min
	SuperScript™ IV	200 U	
	SuperScript™ IV Q190N	200 U	
	SuperScript™ IV Q190F	200 U	
	SuperScript™ IV K103A	200 U	
	SuperScript™ II	200 U	42°C 30 min
	RevertAid™ RT	200 U	
	AMV RT (NEB)	10 U	
	MarathonRT (Kerafast)	200 U	
	HIV RT (Cambio)	30 U	37°C 30 min
RNAP	T7 RNAP V783M	200 U	37°C 3 h
	PUP (NEB)	2 U	37°C 10 min

Reaction products were resolved on 15% TBE-Urea PAGE. Prior to loading on a gel, samples were mixed in a 1:1 ratio with 2× DNA loading buffer (98% formamide, 10 mg/mL blue dextran, 10 mM EDTA), heated at

95°C for 5 min and then immediately cooled on ice. Electrophoresis was carried out in 1× TBE buffer at 400 V for 1 h at 55°C. Gels were imaged with Typhoon™ FLA 9500 system (GE Healthcare).

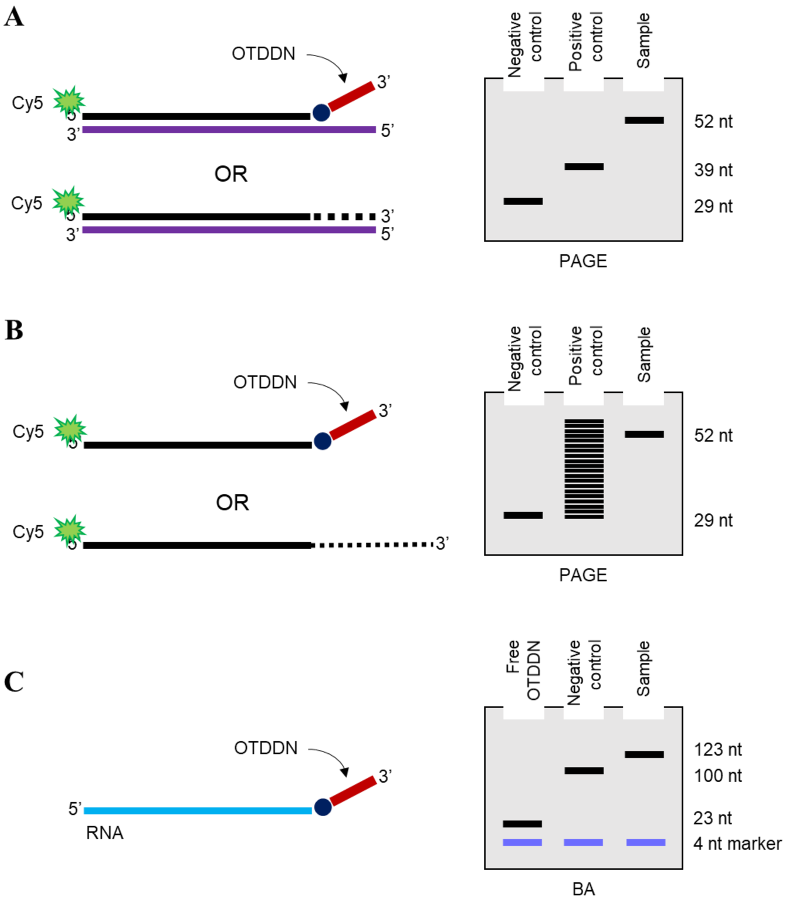


Figure 2.2. OTDDN incorporation testing assays used in this study. In all cases unprocessed primers were used as negative controls, while reaction products after incorporation of unmodified nucleotides – as positive controls. **(A)** – PEX assay employed for testing of DNA polymerases. **(B)** – tailing assay used for TdT testing. **(C)** – tailing assay used for PUP testing. BA – Agilent 2100 Bioanalyzer system.

Testing of RNA polymerases. RNA polymerases were examined in *in vitro* transcription or tailing reactions, following protocols recommended by manufacturers (TranscriptAid™ T7 High Yield Transcription Kit, Thermo Scientific, and Poly(U) Polymerase, Cat. No. M0337S, NEB), except that reaction mixtures were supplemented with OTDDN. Reaction products were purified using Agencourt AMPure XP beads (Beckman Coulter) following a standard PCR purification protocol, except that for the binding step 2× sample volume of beads and the equal amount of 96% ethanol were added, and binding step was prolonged for 15 min at room temperature. Moreover, elution was performed at 65 °C for 5 min in nuclease-free water. Products were analyzed with Agilent 2100 Bioanalyzer system using either RNA 6000 Pico kit or Small RNA kit (Agilent Technologies).

The schematic depiction of OTDDN incorporation testing assays is given in Fig. 2.2.

Template for read-through testing. To assess which polymerases are able to read through an unnatural linker within the OTDDN, incorporation product was used as a template for PEX.

DNA primer was labeled by OTDDN upon incorporation by SuperScript IV enzyme. The primer (5′-TGCAGACATGGGTAGGCATCCTTGGCGTA -3′) was first annealed with RNA template oligonucleotide (5′-auacgccaaggaugccuacccaugucugca-3′) in a 1× annealing buffer. The duplex contained a single 5′-A overhang which was used as a template for the incorporation of ddU^{ON2}TP. Reaction conditions were as described above. After incorporation by RT, RNA template was degraded by RNase H (Thermo Scientific) treatment, and reaction products along with free OTDDN were purified by affinity capture of biotinylated moieties with Dynabeads™ M-270 Streptavidin (Thermo Scientific) magnetic beads according to the protocol for nucleic acid purification. To remove free OTDDN, purified nucleic acids were resolved on a 4% E-Gel EX (Thermo Scientific) and full-length incorporation product was gel extracted using PureLink™ Quick Gel Extraction Kit (Thermo Scientific). ~100 fmol of purified incorporation product were used as a template for subsequent PEX.

PEX reaction conditions for read-through testing. For successful application of OTDDN for nucleic acid detection and analysis, it is highly desirable to identify a thermostable DNA polymerase able to read through the linker – this would make OTDDN-containing nucleic acids compatible with PCR amplification.

Read-through was tested with several thermostable family B polymerases and several representatives of families A and RT. Tested polymerases and specific conditions are listed in Table 2.4. In all cases, 1 pmol of primer (5′-CAGACGTGTGCTCTTCC-3′) complementary to ON2 modification was used per reaction. PEX was performed in commercial buffers and with optimal amounts of dNTPs for each tested polymerase.

Table 2.4. Polymerases tested for reading through OTDDN linker

Family	Polymerase, amount per reaction		Conditions
A	Klenow fragment (exo-)	5 U	30ºC 30 min
	Thermo Sequenase	40 U	95ºC 1 min → 60ºC 1 min → 72ºC 5 min/10min/ 15 min*
B	Platinum™ SuperFi™	2 U	
	Phusion (exo-)	20 U	
	Phusion U	2 U	
RT	SuperScript™ IV	200 U	50ºC 30 min

*Here, three different extension times were tested.

Free primer was removed by Exo I (Thermo Scientific) treatment. Double-stranded read-through products were purified by ethanol precipitation. Briefly, the volume of each reaction mixture was brought to 180 µL with nuclease-free water, then 18 µL of 3 M sodium acetate (Thermo Scientific), 4 µL of 5 mg/mL glycogen (Thermo Scientific) and 600 µL of 96% ethanol (Vilniaus degtinė) were added. Precipitation was conducted at -20ºC for 1 h. Next, the samples were centrifuged at 10,000 × g for 30 min at 4ºC. Supernatant was discarded and pellet was washed twice with 200 µL of cold 70% ethanol. Pellet was dried at 37ºC for 5 min and dissolved in nuclease-free water. PEX products were analyzed with Agilent 2100 Bioanalyzer system using Small RNA kit (Agilent Technologies). Annealed oligonucleotide duplex of the exact same sequence and size as anticipated read-through product was used as a positive control which helped to distinguish between full-length read-through product and primer extended only until the linker, while mixtures processed in the absence of enzymes were used as negative controls. Read-through efficiency was calculated as a molar ratio between full-length extension product and the amount of the starting template.

The principle of read-through testing is schematically depicted in Fig. 2.3.

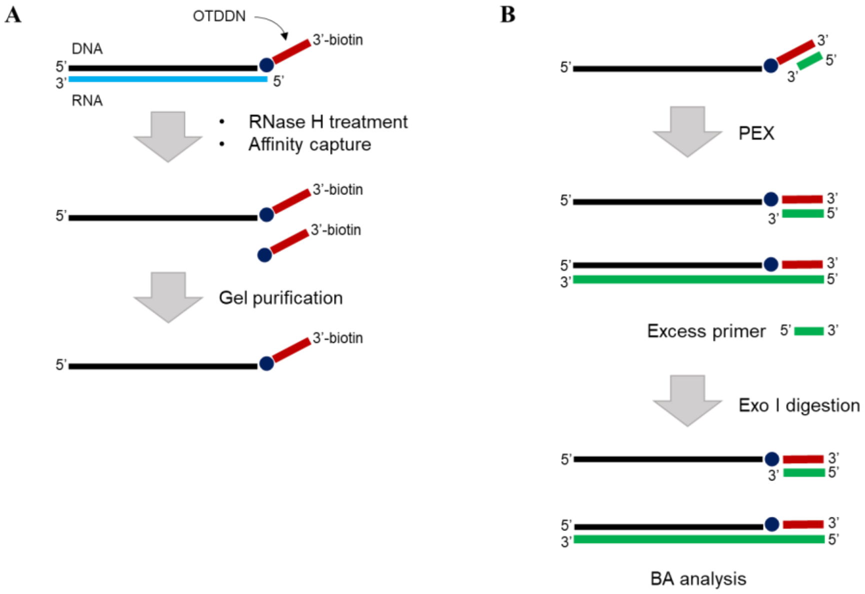


Figure 2.3. The workflow of OTDDN read-through testing. **(A)** – the principle of template preparation. **(B)** – primer extension (PEX) and downstream processing. BA – Agilent 2100 Bioanalyzer system (Small RNA kit was used for the analysis).

2.3 Fragment library generation via *in vitro* transcription

Engineering of T7 RNAP. In certain applications it might be desirable to amplify nucleic acids in linear fashion to avoid biases and errors associated with exponential amplification (Chen *et al.*, 2017). To streamline the sequencing of linear amplification products we sought to engineer T7 RNA polymerase towards the ability to synthesize transcripts in ssDNA form – this would allow direct usage of transcripts as templates in PCR, while OTDDN tagging would simplify fragment library generation from intact (i.e. not linearized) templates. Moreover, enzymatic synthesis of ssDNA would be an attractive platform for the production of oligonucleotides for general molecular biology applications.

To select T7 RNAP variants with reduced substrate specificity, an *in vitro* evolution scheme based on fluorescence-activated droplet sorting was created and validated as described in Kapustina *et al*., 2021b (Fig. 2.4). Briefly, single *Escherichia coli* cells expressing mutant T7 RNAP variants were encapsulated in the presence of lysozyme, *in vitro* transcription (IVT) template and reaction components, where ribopyrimidines were completely substituted by 2′-deoxypyrimidines. Chimeric transcripts were detected in a sequence-specific manner by molecular beacon probes, while all native RNA was digested by

RNase A. Fluorescent droplets were then sorted to retrieve variants with desired activity. From the library of 1463 T7 RNAP mutants, we selected several enzymes able to incorporate 2′-deoxypyrimidines and opportunistically tested them with various combinations of NTPs/dNTPs. Variant V783M exhibited the best performance with dTTP, dCTP, dATP and 2′-F-dGTP substrate mixture.

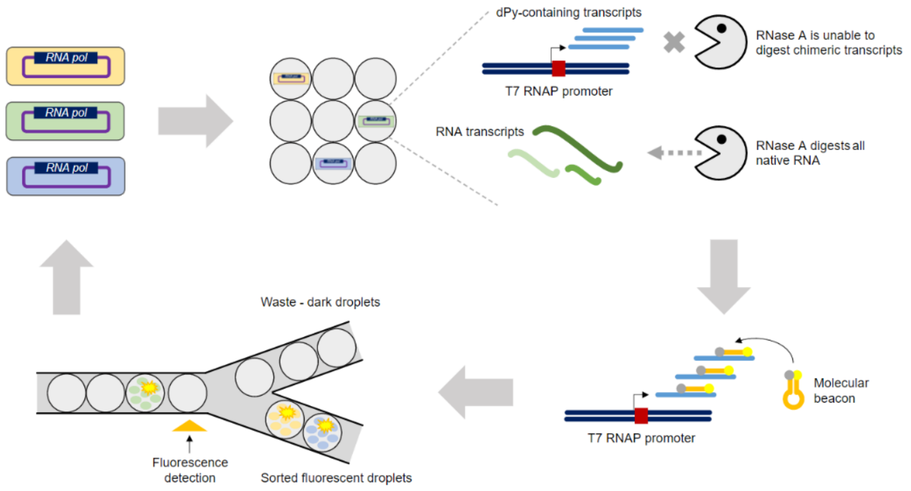

RNA pol
RNA pol
RNA pol
dPy-containing transcripts
T7 RNAP promoter
RNase A is unable to digest chimeric transcripts
RNA transcripts
RNase A digests all native RNA
Waste - dark droplets
Molecular beacon
T7 RNAP promoter
Fluorescence detection
Sorted fluorescent droplets

Figure 2.4. Evolution of T7 RNA polymerase variants able to incorporate deoxynucleotide triphosphates. The workflow is based on the fluorescence-activated droplet sorting technique.

Generation of OTDDN-labeled fragment library via IVT. To test the hypothesis that linear amplification can readily generate tagged fragments that are stable enough to directly serve as PCR templates, IVT was performed in a reaction mixture containing 250 ng of circular pTZ19R plasmid DNA with a segment of Illumina P7 adapter cloned downstream of T7 promoter (P_{T7}), 2 mM of 2′-F-dGTP or GTP, 2.5 mM of dATP, 2.5 mM of dCTP, 2.5 mM of dTTP, 0.5 µM of either ddU^{ON8}TP or ddC^{ON8}TP, 0.1 U of PPase, 8% of DMSO, 200 U of T7 RNAP V783M and 1× TranscriptAid™ reaction buffer. Reaction mixture was incubated at 37°C for 3 h. Transcripts containing 3′-biotinylated OTDDNs were purified using Dynabeads™ M-270 Streptavidin magnetic beads and amplified in a reaction mixture containing 1× Invitrogen™ Collibri™ Library Amplification Master Mix (Thermo Scientific), 20 U of Phusion exo- enzyme, 1 µM of indexing primers from the Invitrogen™ Collibri™ Stranded RNA Library Prep Kit (Thermo Scientific) and nuclease-free water to 50 µL. Cycling conditions were as recommended by the manufacturer. Fragment libraries were then purified using Dynabeads™

Cleanup Beads according to the post-PCR purification protocol from the Invitrogen™ Collibri™ Stranded RNA Library Prep Kit (revision C.0, “Purify the amplified cDNA”). The overview of the experimental scheme is given in Fig. 2.5.

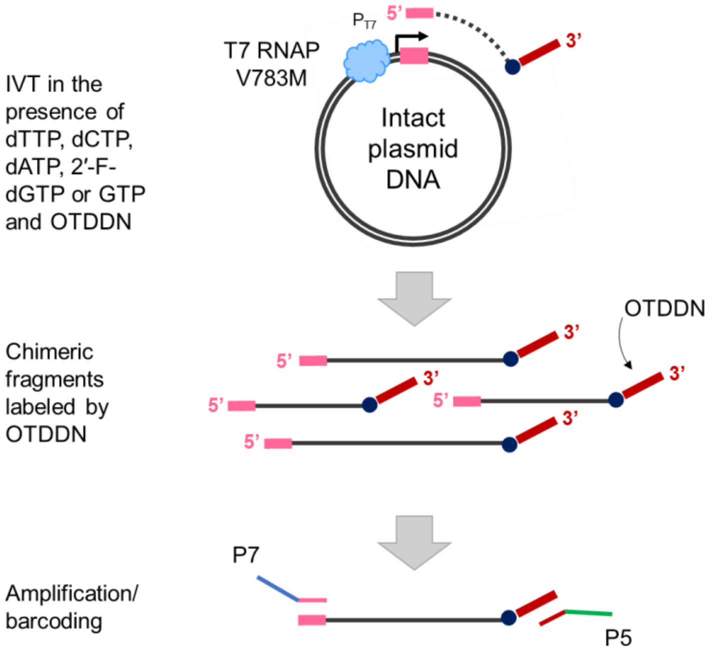

IVT in the presence of dTTP, dCTP, dATP, 2′-F-dGTP or GTP and OTDDN
T7 RNAP V783M
P
T7
5’
3’
Intact plasmid DNA
Chimeric fragments labeled by OTDDN
OTDDN
5’
3’
5’
3’
5’
3’
5’
3’
Amplification/ barcoding
P7
P5

Figure 2.5. The experimental scheme of linear DNA amplification and simultaneous generation of labeled fragments suitable for sequencing by *in vitro* transcription with T7 RNA polymerase mutant V783M.

2.4 DNA sequencing applications

Oligonucleotide-modified chain terminators hold the potential to greatly improve sample preparation for NGS because of their dual functionality: (i) stochastic incorporation of dideoxynucleotides can replace fragmentation step and (ii) simultaneous labeling of corresponding DNA strand with a pre-designed oligonucleotide can replace adapter addition via ligation (Medžiūnė *et al*., submitted). This notion was extensively tested on all major DNA and RNA sequencing applications, making use of identified polymerases able to incorporate OTDDNs and perform read-through.

In DNA sequencing field, any application which is (or may be) based on primer extension reaction, can be re-designed to be compatible with OTDDN technology.

2.4.1 Whole genome library preparation

NGS libraries covering whole genomes can be prepared by employing random priming. For proof-of-principle demonstration, *Escherichia coli* genomic DNA (Thermo Scientific) was used as a template. 100 ng of DNA were used for primer extension reaction with 1 pmol, 10 pmol or 100 pmol of anchored random primers of sequence 5′-TACACGACGCTCTTCCGATCT $(N)_{10}$-3′, 20 pmol of dNTP (each), 2 pmol or 0.2 pmol of $ddU^{ON2}TP$ and 40 U of Thermo Sequenase with TAP in 1× Thermo Sequenase Reaction buffer (Thermo Scientific), reaction volume was 20 µL. Here, the use of thermostable enzyme allows to perform several cycles of linear primer extension with termination by OTDDN. Primer extension reaction was executed as follows: denaturation at 92ºC for 3 min followed by cooling to 16ºC with subsequent incubation for 5 min, slow (+0.1ºC/s) temperature rise to 68ºC with incubation for 15 min, and 15 cycles of denaturation at 92ºC for 30 s, annealing/extension at 68ºC for 5 min, and final extension at 68ºC for 30 min. Half of the primer extension reaction was used directly for indexing PCR which introduced full-length Illumina™ adapters into resulting fragment library. Primer extension reaction mixture was supplemented with 25 µL of 2× Invitrogen™ Collibri™ Library Amplification Master Mix (Thermo Scientific), 20 U of Phusion exo- enzyme (Thermo Scientific), 50 pmol of each of the unique dual indexing primers (see below) and nuclease-free water to 50 µL.

i5 primer: 5′-AATGATACGGCGACCACCGAGATCTACAC[8 nt index] ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′

i7 primer: 5′-CAAGCAGAAGACGGCATACGAGAT[8 nt index]GTGAC TGGAGTTCAGACGTGTGCTCTTCCGATCT-3′

Cycling was performed as follows: denaturation at 98°C for 30 s, followed by 30 cycles of denaturation at 98°C for 10 s, annealing at 60°C for 30 s, extension at 72°C for 1 min, and final extension at 72°C for 1 min. Each PCR reaction was then purified using Dynabeads™ Cleanup Beads (Thermo Scientific). DNA binding to the beads was performed by mixing 65 µL of bead suspension with 50 µL of sample and subsequent incubation at room temperature for 5 min. Sample was then placed on magnet, supernatant was removed and beads were resuspended in 50 µL of elution buffer containing 10 mM Tris-HCl (pH 8.0). 75 µL of fresh beads were added again to the sample and binding was repeated. After room temperature incubation, sample was placed on magnet, supernatant was removed, and beads were washed twice with 85% ethanol. To elute libraries, beads were resuspended in 20 µL of elution buffer and incubated for 1 min at room temperature.

2.4.2 Semi-targeted NGS library preparation from low and high complexity templates

To sequence only specific regions within the template, target-specific primers might be used instead of random primers. The possibility to use only a single primer to capture the region of interest enables so-called semi-targeted design, i.e. one is able to capture known sequence and in addition *a priori* unknown region nearby.

Library preparation from low complexity template. For proof-of-principle demonstration, low complexity 6407 base single-stranded genome of M13mp18 bacteriophage (Thermo Scientific) was used as a sample input. Specific primers were designed to contain partial Illumina P5 adapter anchor sequence and target-specific sequence (Table 2.5). Both primers were oriented in the same direction. The experimental scheme is given in Fig. 2.6.

Table 2.5. Target-specific primers used for library prep from M13mp18 DNA

Name	Oligonucleotide sequence
M13-1	5′-TACACGACGCTCTTCCGATCTAACGGTACGCCAGAATCTTG-3′
M13-2	5′-TACACGACGCTCTTCCGATCTAGAGCCACCACCGGAAC-3′

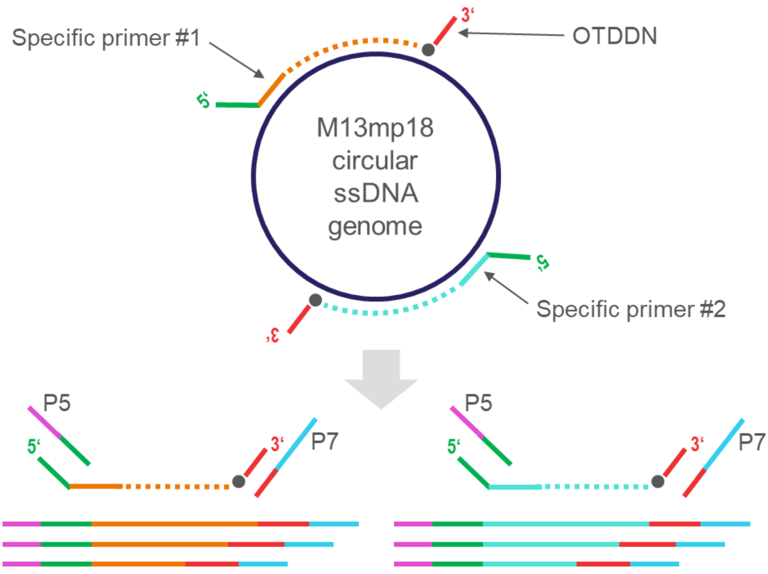


Figure 2.6. The experimental scheme of semi-targeted library preparation from M13mp18 DNA by 2-plex primer extension and tagging of corresponding extension products by OTDDN.

Briefly, 0.125 pmol of each of the primers were mixed with 200 ng of M13mp18 DNA in a reaction mixture containing 2 pmol of ddU^{ON2}TP, 20 pmol of dNTPs (each) and 40 U of Thermo Sequenase in 1× Thermo Sequenase Reaction buffer. Primer extension was executed as follows: denaturation at 95°C for 30 s, followed by 15 cycles of denaturation at 95°C for 30 s, annealing/extension at 60°C for 2 min and final extension at 60°C for 30 min. Half of the reaction was used directly for indexing PCR as described above. Final libraries were purified using Dynabeads™ Cleanup Beads.

Library preparation from high complexity template. Semi-targeted design was further tested on high complexity human genomic DNA (Thermo Scientific). Five primers complementary to specific targets within *ALK* gene were designed for this experiment (Table 2.6).

Table 2.6. Target-specific primers used for library prep from human gDNA

Name	Oligonucleotide sequence
ALK-1	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTGTAGTTGGGGTTGTAGTCGGTCATGATG-3′
ALK-2	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTTTATATAGGGCAGAGTCATGTTAGTCTGG-3′
ALK-3	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTGTGTTTCCTATAGTTGGAGAACTGCCAAG-3′
ALK-4	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTCATTATCACTCCTACATGTGAGGATGTTCG-3′
ALK-5	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTCTTGGCTCACAGGCTGAACAGAAATATAC-3′

Each primer was tested individually. Primer extension reactions were performed with 50 ng or 500 ng of human gDNA, 500 pmol of dNTPs (each), 5 pmol of ddU^{ON2}TP, 2 pmol of ddC^{ON2}TP, 10 pmol of primer and 40 U of Thermo Sequenase in 1× Thermo Sequenase reaction buffer. Here, two different OTDDNs were used to increase sequence diversity at the position of OTDDN incorporation – this is important to ensure base calling of good quality (Kircher *et al.*, 2011). Cycling conditions were as follows: denaturation at 95°C for 4 min, followed by 15 cycles of denaturation at 95°C for 1 min, annealing at 60°C-67°C (depending on the melting temperature of each individual primer), extension at 72°C for 1 min, and final extension at 72°C for 5 min.

To remove residual oligonucleotides and template, primer extension products were enriched utilizing biotin modification within OTDDN. Reaction products were purified with Dynabeads™ M-270 Streptavidin magnetic beads and then subjected to indexing PCR as described above, except that the number of PCR cycles was 20. After post-PCR cleanup, libraries were reamplified using Invitrogen™ Collibri™ Library Amplification Master Mix with Primer Mix (Thermo Scientific), the number of PCR cycles was 12. Such reamplification is needed to achieve sufficient library yield for Illumina sequencing as capturing single region within high complexity template results in very limited amounts of target molecules. After reamplification samples were purified using Dynabeads™ Cleanup Beads as described above, except that for the first binding 45 μL of beads were used, and for the second bind step – 50 μL.

2.4.3 Development of a new approach for characterization of bacterial communities

After semi-targeted DNA library preparation proved to be feasible, we sought to apply it to solve some inherent issues of amplicon sequencing. The capture of unknown genomic region adjacent to the bacterial 16S rRNA gene can help to determine gene copy number and achieve higher taxonomic resolution which would make the characterization of bacterial communities much more accurate. This assumption was firstly checked by *in silico* modeling. Next, the workflow for semi-targeted bacterial sequencing was developed and benchmarked against currently available solutions (Kapustina *et al.,* 2021a).

Samples. For proof-of-principles experiments, genomic DNA extracted from *Escherichia coli* BL21 cells was used as a sample. Cells were cultured from a single colony, and DNA was extracted using GeneJET™ Genomic DNA Purification Kit (Thermo Scientific).

The performance of the developed workflow was then tested on commercially available microbial community standards: ZymoBIOMICS™ Microbial Community DNA Standard (Zymo Research, Cat. No. D6305) and ATCC 20 Strain Even Mix Genomic Material (ATCC® MSA-1002™).

To test the workflow on real metagenomic samples, DNA was extracted from 250 mg of soil using ZymoBIOMICS™ DNA Miniprep Kit (Zymo Research). Soil microbial communities contain the highest level of prokaryotic diversity of any environment (Delmont *et al*. 2011) making it an interesting yet challenging sample type for DNA metabarcode sequencing studies.

Semi-targeted 16S rRNA sequencing. Primers complementary to the conservative 16S rRNA gene site between V2 and V3 variable regions and oriented towards the upstream of the gene were designed (Table 2.7). Sequences were selected on the basis of SILVA (Quast *et al*., 2013) release 132 dataset “SSU Ref NR 99” in a way which maximizes sensitivity toward bacterial rRNA genes ensuring that the annealing temperature would not drop below a certain threshold due to mismatches.

Table 2.7. Target-specific primers used for semi-targeted 16S rRNA sequencing

Name	Oligonucleotide sequence
PR-1	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTTCCCCACTGCTGCCTCCCGTAGGAG-3′
PR-2	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTACGCGGCGTCGCTGCATCAGG-3′
PR-3	5′-CTCTTTCCCTACACGACGCTCTTCCGATCTGCAAGATTCCCCACTGCTGCCTCCCGTAGG-3′

Library preparation from 30 ng of metagenomic DNA was based on the extension of an equimolar primer mix using Thermo Sequenase enzyme and OTDDNs bearing ON2 modification as described in Kapustina *et al.*, 2021a (see Methods). Tagged fragments were then enriched using Dynabeads™ M-270 Streptavidin magnetic beads and amplified. The schematic representation of the workflow is provided in Fig. 2.7.

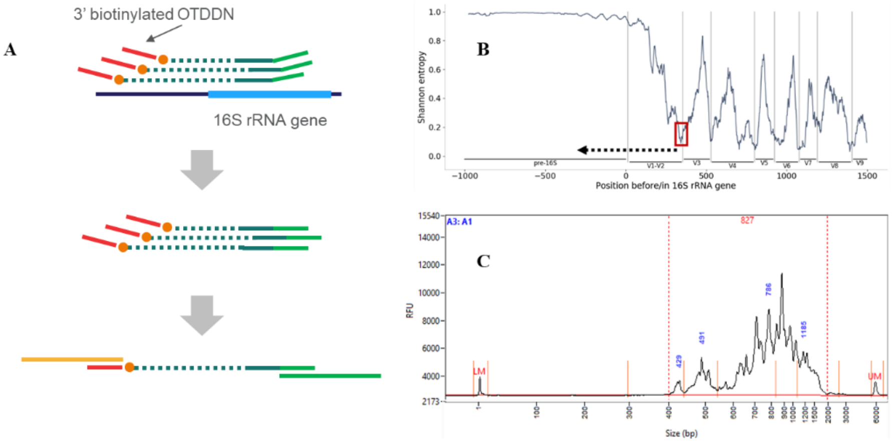


Figure 2.7. The experimental scheme of semi-targeted library preparation from metagenomic DNA. **(A)** – library preparation starts from the extension of primers specific to 16S rRNA gene; extension products are then enriched and amplified introducing full-length sequencing adapters. **(B)** – primers are designed to target conservative part of the 16S rRNA gene (marked in red). Primer extension captures 16S rRNA V1-V2 regions and also extends into genomic sequences which are more diverse than the sequence of 16S rRNA gene. **(C)** – an example of semi-targeted library electropherogram.

Whole metagenome sequencing. Unbiased characterization of microbial composition of metagenomes was conducted by WGS. Briefly, 20 ng of DNA in 50 µL volume of 10 mM Tris-HCl (pH 8) were sheared with Covaris™ E220 Evolution Focused-ultrasonicator applying the following conditions: peak incident power – 175 W, duty factor – 10%, cycles per burst – 200, treatment time – 50 s. Such conditions result in DNA fragment size distribution median at ~300 bp.

25 µL of fragmented DNA solution were then used for library preparation using the Invitrogen™ Collibri™ PS DNA Library Prep Kit according to the manufacturer′s recommendations.

Benchmarking against conventional 16S rRNA sequencing techniques. To compare the performance of semi-targeted method with conventional approaches, benchmarking study was conducted. Libraries were prepared from microbial community standards and soil DNA with the kits listed in the Table 2.8. DNA input amount was within the recommended range for each individual kit (1-10 ng). All libraries were prepared with two technical replicates.

Table 2.8. Commercially available NGS library preparation kits for high-throughput 16S rRNA gene sequencing selected for comparative analysis

Kit/Protocol	Cat. No.	Supplier	Analyzed region
QIAseq™ 16S/ITS Screening Panel	333812	Qiagen	Whole gene
Swift™ Amplicon 16S+ITS Panel	AL-51648	Swift Biosciences	Whole gene
NEXTFLEX™ 16S V1-V3 Amplicon-Seq Kit	NOVA-4202-02	PerkinElmer	V1-V3
NEXTFLEX™ 16S V4 Amplicon-Seq Kit 2.0	NOVA-4203-02	PerkinElmer	V4
Quick-16S™ NGS Library Prep Kit	D6410	Zymo Research	V1-V2
16S Illumina amplicon protocol	N/A	EMP	V4

2.5 RNA sequencing applications

To sequence transcriptomes on a short read sequencer, RNA should be converted into cDNA fragment library flanked by the appropriate adapters. This makes RNA library preparation a more technically complex process than DNA library prep, with additional challenges related to the capture of RNA molecules of interest and to retaining the accurate quantitative representation.

OTDDNs were tested in whole transcriptome and gene expression applications, both in bulk and single-cell levels.

2.5.1 PCR-free cDNA fragment library preparation

Direct sequencing of first-strand cDNA without PCR amplification is not possible with conventional library preparation approaches because there is no convenient way to add an adapter on a 3′ terminus of a cDNA fragment. OTDDNs may hold the solution, however sequencing chemistry should accept OTDDN-containing DNA as a template for bridge amplification on the surface of a flow cell.

To test the tolerance of sequencing chemistry to an unnatural linker within OTDDN, a model system was created. 10 pmol of oligonucleotide containing full-length P5 Illumina adapter and a short sequence which served as a sequencable insert (P5-miR) was labeled with OTDDN bearing long ON10 modification by TdT treatment in a reaction mixture containing 50 pmol of ddC^{ON10}TP and 40 U of TdT in 1× TdT reaction buffer. Tailing reaction was performed at 37ºC for 1 h and then stopped by the addition of EDTA to 50 mM final concentration. Reaction products were purified using Dynabeads™ M-270 Streptavidin beads. Successful tailing was confirmed by capillary electrophoresis using Agilent 2100 Bioanalyzer Small RNA Kit (Fig. 2.8).

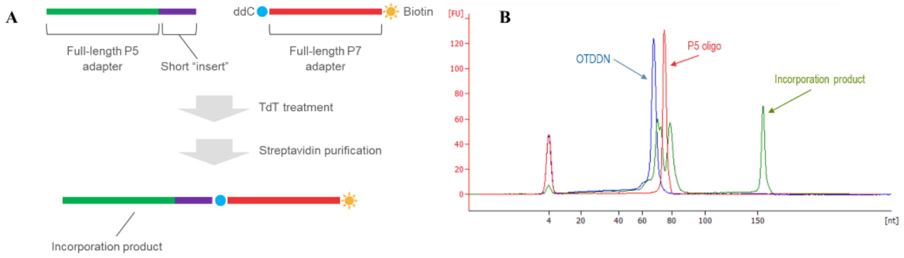


Figure 2.8. (**A**) **-** the depiction of a control system used to test the tolerance of Illumina sequencing chemistry to OTDDN linker. (**B**) – electropherogram showing oligonucleotides used for tailing and tailing product.

To account for any effects which may arise during sequencing because of the presence of an unnatural OTDDN linker, control fragment with exactly the same insert was prepared. Oligonucleotide P5-miR-biot (50 pmol) was annealed with either P7-ix2 (55 pmol) or P7-ix4 (55 pmol) in Klenow buffer. Then, reaction mixture was supplemented with dNTPs (0.2 mM final concentration of each) and 5 U of Klenow exo-. Oligonucleotide extension was performed at 30ºC for 30 min. The resulting duplex was immobilized on Dynabeads™ M-270 Streptavidin beads and non-biotinylated DNA strand was dissociated by NaOH treatment according to the manufacturer′s instructions.

Biotinylated strand, which mimics the OTDDN tailing product, was then eluted. Tailing product and two control fragments were pooled and sequenced.

Oligonucleotides used in this experiment are listed in Table 2.9.

Table 2.9. Oligonucleotides used to design a control system for direct sequencing of OTDDN-tagged ssDNA. Sequences which correspond to the insert are underlined.

Name	Oligonucleotide sequence
P5-miR	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAACCACACAACCTACTACCTCA-3′
P5-miR-biot	5′-biotin-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAACCACACAACCTACTACCTCA-3′
P7-miR-ix2	5′-CAAGCAGAAGACGGCATACGAGATTCAGATTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGATGAGGTAGTAGGTTGTGTGGTT-3′
P7-miR-ix4	5′-CAAGCAGAAGACGGCATACGAGATTCGAAGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGATGAGGTAGTAGGTTGTGTGGTT-3′

Once compatibility with Illumina sequencing chemistry was confirmed, the generation of PCR-free cDNA libraries was attempted. 500 ng of universal human reference RNA (UHRR; Thermo Scientific) were reverse transcribed in a following reaction mixture: 50 pmol of RT primer (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT(T)$_{30}$-3′), 20 pmol of dNTP (each), either 20 pmol or 2 pmol of ddC^{ON10}TP, 5 mM DTT and 200 U of SuperScript IV in 1× RT buffer. RT was performed at 50ºC for 30 min, reverse transcriptase was then inactivated by heating at 80ºC for 10 min. To remove template RNA, RT reaction mixture was treated with RNase H at 37ºC for 20 min. Single-stranded cDNA fragments labeled by OTDDN were purified with Dynabeads™ M-270 Streptavidin beads and sequenced. During template preparation for MiSeq™, NaOH denaturation step was omitted.

2.5.2 Whole transcriptome library preparation

Similarly to WGS, OTDDN labeling might be applied for random primed cDNA library prep covering the whole transcript length. Whole transcript coverage enables not only quantitative gene expression measurements but also structural studies of splice isoforms.

To capture only the coding portion of the transcriptome, mRNA was first converted into cDNA, and during the second strand synthesis step tagged library fragments were generated (Fig. 2.9). 1 μg of UHRR were reverse transcribed in a following reaction mixture: 50 pmol of RT primer (5′-AAGCAGTGGTATCAACGCAGAGTAC(T)$_{30}$-3′), 0.5 mM dNTP (each), 5 mM DTT, 200 U of SuperScript IV in 1× RT buffer. Reaction was performed at 50ºC for 10 min and terminated by heating at 80ºC for 10 min. RNA templates were digested by RNase H at 37ºC for 20 min. Reaction products were purified with Dynabeads™ Cleanup Beads according to the protocol described in Collibri™ Stranded RNA Library Prep Kit user guide (revision C.0, section „Purify the fragmented RNA").

Purified cDNA was transferred to the second strand synthesis reaction containing 5-0.005 µM of anchored random primers (5´- TACACGACGCTC TTCCGATCT$(N)_{10}$-3´), 2 pmol of each of ddU^{ON2}TP and ddC^{ON2}TP, 20 pmol of dNTP (each) and 40 U of Thermo Sequenase enzyme in 1× Thermo Sequenase reaction buffer. Primer annealing and extension was performed as described in section 2.3.1, except that here a single extension cycle was performed. Reaction products were subjected to indexing PCR as described in 2.3.1, except that the number of PCR cycles was 20.

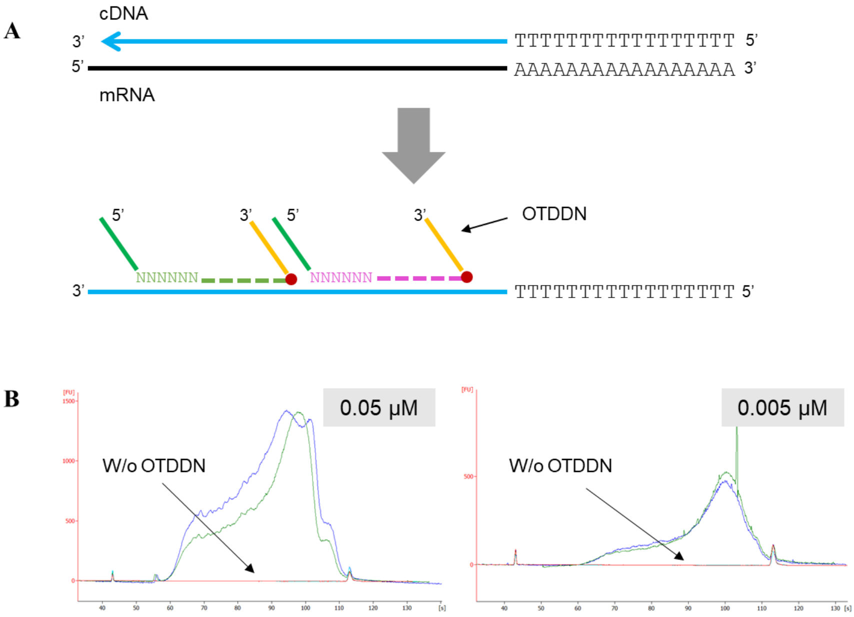


Figure 2.9. (**A**) – the principle of whole transcriptome library preparation with OTDDNs, focusing specifically on protein coding RNAs. (**B**) – typical electropherograms of whole transcriptome libraries. Green and blue traces correspond to technical replicates for each condition. The number above the traces stands for the concentration of random primers used in second strand synthesis reaction.

mRNA 5´end enrichment. To capture protein coding transcripts and specifically enrich for 5′ termini, 100 ng – 1 µg of UHRR were reverse transcribed in a following reaction mixture: 50 pmol of oligo(dT)$_{30}$ RT primer, 0.5 mM dNTP (each), 5 mM DTT, 200 U of SuperScript IV in 1× RT buffer. Reaction was performed at 50ºC for 10 min, then 20 pmol of template switching oligonucleotide (TSO) of sequence 5′-CCAGGACCAGCGATTC ggg-3′ were added to the reaction and reverse transcription was continued at 50ºC for 15 min. Afterwards, nucleic acids were purified with Dynabeads™ Cleanup Beads following the Collibri™ Stranded RNA Library Prep Kit user guide (revision C.0, section „Purify the fragmented RNA"). To synthesize second strand tagged by OTDDN, cDNA was transferred to the following reaction mixture: 1 pmol of second strand synthesis primer (5′-CAGTGGTATCAACGCAGAGTACCCAGGACCAGCGATTC-3′), 2 pmol of ddU^{ON2}TP, 20 pmol of each of dNTP and 40 U of Thermo Sequenase in 1× Thermo Sequenase Reaction buffer. Second strand synthesis reaction was executed as follows: denaturation at 95ºC for 3 min, 10 cycles of denaturation at 95ºC for 30 s and annealing/extension at 60ºC for 2 min, and a final extension at 60ºC for 5 min. Reaction products were purified using Dynabeads™ M-270 Streptavidin beads and subjected to PCR. Amplification conditions were as described in 2.3.1, except that the i5 primer was replaced with the following oligonucleotide: 5′- AATGATACGGCGACCACCGAGA TCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC-3′, and the number of PCR cycles was 25. The final library was purified using Dynabeads™ Cleanup Beads.

The principle of the described workflow is summarized in Fig. 2.10.

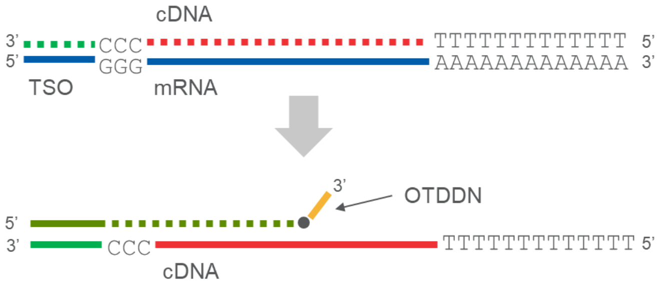


Figure 2.10. The principle of library construction with OTDDNs targeting 5′ termini of mRNA transcripts.

mRNA 3′end enrichment. The enrichment for 3′ termini is technically less complex than that for 5′ termini and is widely adopted for single-cell gene expression profiling, thus 3′-end sequencing was explored in more detail.

Proof-of-principle was demonstrated using well-characterized human total RNA – UHRR and HeLa – as sample input. Libraries were prepared from 100 pg – 1 µg of RNA and spiked with Invitrogen™ ERCC ExFold Spike-In Mixes (Thermo Scientific). In addition, direct library construction from eukaryotic cell lysates was tested using 10 – 10 000 HEK-293 or BALB/3T3 cells suspended in 1× PBS, pH 7.4. To compare the library prep performance from cell lysates and from purified RNA, total RNA was extracted from 1 M HEK-293 cells using the Invitrogen™ PureLink™ RNA Mini Kit according to the manufacturer′s instructions.

Reverse transcription was performed in 20 µL reaction mixture containing 200 U of SuperScript IV reverse transcriptase, 50 pmol of RT primer of sequence 5′-CTGGAGTTCAGACGTGTGCTCTTCCGATCT(T)$_{30}$-3′, 20 pmol of dNTP mix, 40 U of RiboLock RNase Inhibitor, 5 mM DTT, 2 pmol of ddU^{ON9}TP, 0.4 pmol of ddC^{ON9}TP in 1× of SuperScript IV RT buffer. For library preparation from cell lysates, reverse transcription reaction was supplemented with 0.3% IGEPAL™ CA-630 (Sigma-Aldrich) to ensure cell lysis. Reaction was performed for 30 min at 50°C followed by termination at 80°C for 10 min. After reverse transcription, reaction mixture was used directly for cDNA amplification. Amplification conditions were as described in 2.3.1., the number of PCR cycles was 10-25 depending on the amount of starting material.

Each PCR reaction was then purified using Dynabeads™ Cleanup Beads. DNA binding to the beads was performed by mixing 45 µL of bead suspension with 50 µL of sample and subsequent incubation at room temperature for 5 min. Sample was then placed on magnet, supernatant was removed and beads were resuspended in 50 µL of elution buffer containing 10 mM Tris-HCl (pH 8.0). 50 µL of fresh beads were added again to the sample and binding was repeated. After room temperature incubation, sample was placed on magnet, supernatant was removed, and beads were washed twice with 85% ethanol. To elute libraries, beads were resuspended in 15 µL of elution buffer and incubated for 1 min at room temperature.

To generate enough material for sequencing, low RNA inputs (100 pg – 500 pg) required an additional amplification step. Reamplification was performed in a 50 µL reaction with Invitrogen™ Collibri™ Library Amplification Master Mix with Primer Mix for 6-12 cycles according to the recommended temperature conditions. Final libraries were purified using Dynabeads™ Cleanup Beads as described above.

The overview of the library construction principle is shown in Fig. 2.11.

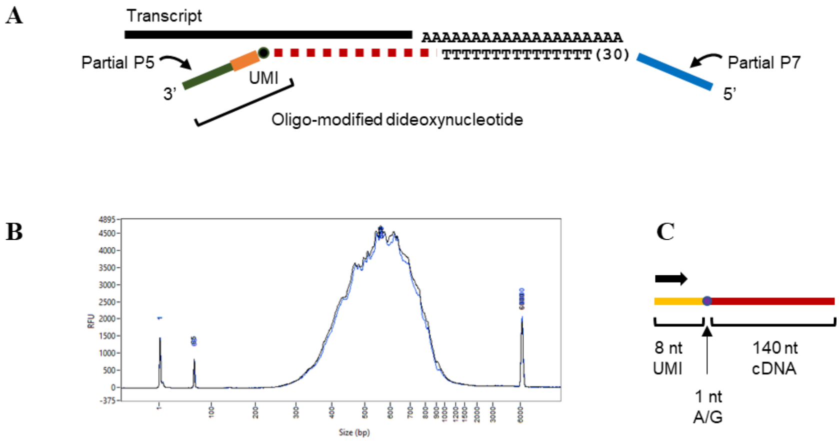


Figure 2.11. (**A**) **–** the principle of library construction employing OTDDNs and targeting 3′ termini of mRNA transcripts. (**B**) – an example of a typical library trace. Blue and black traces correspond to technical replicates. (**C**) – the structure of the sequencing read: first 8 nt of Read 1 correspond to the in-line UMI, the next nucleotide corresponds to the ddNTP base within OTDDN, and afterwards a cDNA fragment is being sequenced.

Benchmarking against conventional 3´end sequencing techniques. To assess the performance of the developed gene expression profiling technique, it was compared to commercially available Collibri™ 3′ mRNA Library Prep Kit for Illumina™ Systems (Thermo Scientific). Libraries were prepared from 500 pg, 10 ng, 100 ng and 500 ng UHRR and HeLa total RNA, with ERCC ExFold mixes. Using commercial kit, libraries were prepared with strict adherence to manufacturer′s instructions. All samples were processed in triplicates.

2.5.4 Single-cell RNA sequencing

Sequencing transcriptomes of thousands of cells at single-cell level resolution became possible once high-throughput cell isolation techniques emerged. From the technological perspective, apart from cell isolation the chemistry behind library construction is very similar to that of bulk sequencing methods. We reasoned that OTDDN technology can be applied for labeling of cellular cDNAs captured on barcoding beads. To analyze this possibility, we sequenced Drop-seq libraries prepared from mixtures of HEK-293 and BALB/3T3 cells and compared technical data parameters with the dataset obtained from same cell mixtures through OTDDN-based Drop-seq workflow.

Conventional Drop-seq workflow. In 2015 Macosko *et al.* published the Drop-seq technique developed using in-house microfluidic devices (Fig. 2.12). Briefly, cells are encapsulated in microdroplets with beads coated with barcoded RT primers, and lysis buffer. Upon cell lysis, released mRNA hybridizes to the beads. The emulsion is then broken, and beads are subjected to RT with template switching. Obtained STAMPs (single-cell transcriptomes attached to microparticles) are used as templates for transcriptome preamplification PCR. Subsequently, amplification reaction products are used as input for transposase-based library preparation methods, such as Nextera™ XT DNA Library Prep Kit (Illumina).

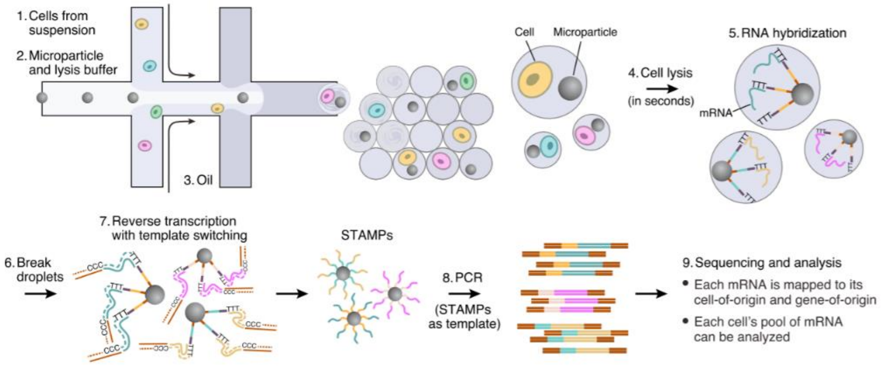

1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
Cell
Microparticle
4. Cell lysis
(in seconds)
5. RNA hybridization
mRNA
6. Break droplets
7. Reverse transcription with template switching
STAMPs
8. PCR
(STAMPs as template)
9. Sequencing and analysis
• Each mRNA is mapped to its cell-of-origin and gene-of-origin
• Each cell's pool of mRNA can be analyzed

Figure 2.12. The overview of single-cell RNA library preparation according to the conventional Drop-seq protocol (by Macosko *et al.*, 2015).

In this study, we co-encapsulated cells with barcoded beads using Dolomite Bio Nadia™ instrument (Blacktrace Holdings Ltd). Barcoded beads were obtained from ChemGenes (Cat. No. MACOSKO-2011-10). Preparation of cells, encapsulation conditions and subsequent library preparation steps were executed as described in Dolomite Bio user guide “DropSeq on the Nadia Instrument”, version 1.2, with the following exceptions: (i) Maxima™ H Minus Reverse Transcriptase and its reaction buffer (Thermo Scientific) were replaced with SuperScript IV RT, and the incubation step at 42°C for 90 min during reverse transcription was replaced with incubation at 50°C for 60 min; (ii) tagmentation of amplified cDNA was performed using the MuSeek™ Library Preparation Kit, Illumina compatible (Thermo Scientific).

Tagmentation protocol was optimized to be compatible with low DNA inputs. MuSeek Enzyme Mix was diluted 100-fold with MuA dilution buffer

(Thermo Scientific) prior to use. 600 pg of amplified cDNA were combined with 0.9 µL of diluted MuSeek Enzyme Mix and MuSeek Fragmentation Reaction Buffer in a total volume of 12 µL. Tagmentation was conducted at 30°C for 30 min. Reaction was terminated by the addition of 1.2 µL of MuSeek Stop Solution. Tagmented DNA was purified using Dynabeads™ Cleanup Beads as described in 2.3.1, except that only one round of binding was performed using 40 µL of magnetic beads. Tagged fragments corresponding to barcoded 3′ termini were then amplified as specified in the user guide for MuSeek™ Library Preparation Kit, except that MuSeek indexed primer M5XX was replaced with the oligonucleotide of sequence 5′-AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGTATCAACGCAGAGTAC-3′. Libraries were sequenced using custom primers listed in Table 2.10.

Table 2.10. Primers used to sequence conventional Drop-seq libraries

Name	Oligonucleotide sequence
Read 1	5′-GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC-3′
Index Read	5′-TGAACTGACGCACGAACTGCACTCGACCTCG-3′
Read 2	5′-CGAGGTCGAGTGCAGTTCGTGCGTCAGTTCA-3′

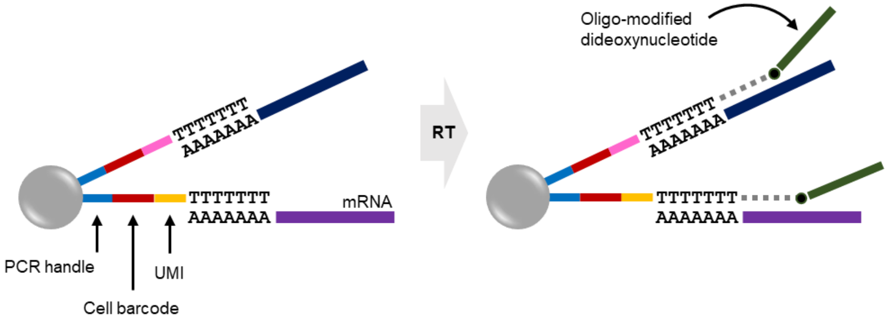


Figure 2.13. Adaptation of OTDDN technology for single-cell sequencing. cDNA labeling eliminated the need for template switching and transcriptome preamplification.

Modified Drop-seq workflow. cDNA fragmentation and labeling by OTDDN eliminates the need for template switching and preamplification steps (Fig. 2.13). This results in nearly twice as fast protocol as compared to the conventional technique. The modified protocol includes an additional second strand synthesis step before Exo I treatment – this is necessary to avoid degradation of tagged cDNAs through the hydrolysis of ddNTP. The protocol does not include tagmentation library prep because sequencing-ready libraries are obtained after the amplification of STAMPs.

Cell mixture was co-encapsulated with barcoded beads as described in Dolomite Bio user guide "DropSeq on the Nadia Instrument", version 1.2. After emulsion breakage and bead wash steps, beads were resuspended in 200 µL of RT reaction mix containing 1× SuperScript IV RT buffer, 20 pmol of ddC^{ON2}TP, 200 pmol of dNTP mix, 5 mM DTT, 4% Ficoll PM-400, 1 U/µL RiboLock™ RNase Inhibitor, and 10 U/µL SuperScript IV reverse transcriptase. Beads were incubated at 25°C for 30 min and then at 50°C for 1 hour in a thermomixer with shaking at 1400 rpm. Afterwards, beads were collected by centrifugation at 1000× g for 1 min, reaction mixture was removed, and beads were washed with 1 mL of TE/SDS (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.5% SDS) three times, then with 1 mL of TE/TW solution (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.01% Tween™-20) once, and finally with 1 mL of nuclease-free water. For the second strand synthesis step, beads were resuspended in 200 µL of reaction mixture containing 1× Phusion™ GC buffer, 1 U/µL Phusion exo- enzyme, 1 µM second strand synthesis primer of sequence 5′-CAGACGTGTGCTCTTCC-3′ and 0.1 mM dNTP mix. Suspension was divided into four PCR tubes in 50 µL aliquots and placed into thermal cycler. Second strand synthesis conditions were as follows: denaturation at 95°C for 1 min, annealing at 60°C for 1 min, extension at 72°C for 15 min. Next, reaction mixtures from a single sample were combined, beads were collected by centrifugation and washed with TE/SDS twice, then with TE/TW once, and finally with 10 mM Tris-HCl (pH 8.0) once. To remove the excess of free primers on beads, they were resuspended in 200 µL of reaction mixture containing 1× Exonuclease I buffer and 1 U/µL Exo I (Thermo Scientific). The suspension was incubated at 37°C for 45 min in a thermomixer with shaking at 1400 rpm. Thereafter, beads were washed with TE/SDS once, twice – with TE/TW, and finally with nuclease-free water once. Beads were resuspended in 200 µL of nuclease-free water.

~2000 beads were used in a single library amplification reaction. Bead concentration was determined using Fuchs-Rosenthal hemocytometer (NanoEnTek), the concentration was adjusted to ~200 beads/µL with nuclease-free water. Library amplification was performed as described in

2.3.1, except that KAPA™ HiFi™ DNA polymerase (Roche) was used. Libraries were purified using Dynabeads™ Cleanup Beads as described for bulk 3′ mRNA libraries.

2.6 NGS library quality control and sequencing

The quality of all sequencing-ready libraries was assessed via capillary electrophoresis either with Agilent 2100 Bioanalyzer High Sensitivity DNA kit or Fragment Analyzer NGS High Sensitivity kit (Agilent Technologies).

The concentration of sequenceable molecules was determined by qPCR using the Invitrogen™ Collibri™ Library Quantification Kit according to the manufacturer´s recommendations.

Sequencing was performed on an Illumina™ MiSeq™ instrument using the following reagent kits:

- MiSeq Reagent Nano/Micro Kits v2 (300-cycle) for pilot and quality control runs or low complexity samples;
- MiSeq Reagent Kit v2 (300-cycle) for routine DNA and RNA sequencing (WGS, whole transcriptome sequencing);
- MiSeq Reagent Kit v3 (150-cycle) for gene expression analysis runs;
- MiSeq Reagent Kit v3 (600-cycle) for 16S rRNA gene sequencing.

Deeper sequencing was performed on an Illumina™ NovaSeq™ 6000 system at Novogene Europe (Cambridge, United Kingdom). Sequencing runs were performed either in paired-end (2×150 bp, 2×300 bp, 21+100 bp) or single-read (1×150 bp) modes depending on the library structure.

Low complexity samples were mixed with 5-20% of PhiX Control v3 library (Illumina) prior to sequencing.

2.7 Data analysis

NGS data analysis included standard pre-processing steps, such as quality filtering, adapter trimming, sequencing depth normalization, with subsequent read alignment to the reference genome and downstream analyses which pipelines depended on the studied library preparation assay. Data analysis was performed by a team of bioinformaticians – dr. Gediminas Alzbutas, dr. Varvara Dubovskaja, Karolis Matjošaitis and Gytis Mackevičius.

3. RESULTS

The results of this thesis constitute three major parts: (I) enzymatic incorporation of OTDDNs and their biocompatibility; (II) utility of OTDDNs for high-throughput DNA sequencing and the development of new microbiome analysis method; (III) utility of OTDDNs for high-throughput RNA sequencing and the development of gene expression analysis technique. The synthesis of OTDDNs used in this work was executed by J. Medžiūnė and will be described in detail elsewhere.

3.1 Enzymatic processing of OTDDNs

A plethora of commercially available and internally developed DNA and RNA polymerases were tested for the ability to process OTDDNs. From a practical perspective, the enzymes desired to incorporate OTDDNs should have been representatives from family RT and thermostable DNA polymerases from either family A or B. A compatible RT enzyme would enable cDNA labeling, while thermostable DNA polymerase would allow to execute workflows with dsDNA templates. In the majority of applications, OTDDN-labeled DNA fragments would be subsequently subjected to PCR amplification, thus DNA synthesis through an unnatural triazole linkage should ideally be carried out by a thermostable DNA polymerase. As will be described in following chapters, enzymes exhibiting appropriate properties were identified.

3.1.1 Substrate properties of OTDDNs

Incorporation assays tested not only enzyme′s ability to accept bulky ON modification attached to the nucleobase, but also the discrimination against ddNTPs. As could be expected, the best candidates among DNA polymerases from families A and B were exonuclease-deficient enzymes with intrinsic or engineered tolerance to nucleotide analogs, such as Thermo Sequenase, CycleSeq and Sequenase V.2.0. Many reverse transcriptases were able to catalyze the incorporation of OTDDN. Moreover, TdT, T7 RNAP variant and PUP were found to accept modified substrate well.

All enzymes were tested with ddU^{ON}TP analogs, and selected candidates (Thermo Sequenase and SuperScript IV RT) were additionally challenged with all other nucleotide analogs. The results of screening are summarized in Table 3.1. The data supporting successful incorporation is provided in Supplementary Fig. 1.

Table 3.1. Incorporation of oligonucleotide-tethered dideoxynucleotides by DNA and RNA polymerases

Polymerase	**OTDDNs**			
	ddU^{ON}TP	**ddC^{ON}TP**	**ddG^{ON}TP**	**ddA^{ON}TP**
Taq	-	nt	nt	nt
Taq (exo-)	-	nt	nt	nt
Tth	-	nt	nt	nt
Platinum™ II *Taq*	-	nt	nt	nt
DyNAzyme™ II	-	nt	nt	nt
DyNAmo™ IV	-	nt	nt	nt
Thermo Sequenase	++	++	++	+
CycleSeq™	++	nt	nt	nt
KlenTaq	-	nt	nt	nt
Sequenase™ V2.0	++	nt	nt	nt
T7 DNA polymerase	+	nt	nt	nt
Klenow fragment (exo-)	-	nt	nt	nt
Bsm DNA polymerase	-	nt	nt	nt
DNA polymerase I	-	nt	nt	nt
Platinum™ SuperFi™	-	nt	nt	nt
Phusion™ (exo-)	+	nt	nt	nt
Pfu (exo-)	-	nt	nt	nt
Phusion™ U	-	nt	nt	nt
T4 DNA polymerase	-	nt	nt	nt
Phi29 polymerase	-	nt	nt	nt
TdT	++	nt	nt	nt
Maxima™ RT	++	nt	nt	nt
SuperScript™ IV	++	++	+	+
SuperScript™ IV Q190N	++	nt	nt	nt
SuperScript™ IV Q190F	-	nt	nt	nt
SuperScript™ IV K103A	+	nt	nt	nt
SuperScript™ II	++	nt	nt	nt
RevertAid™ RT	++	nt	nt	nt
AMV RT	-	nt	nt	nt
MarathonRT	-	nt	nt	nt
HIV RT	++	nt	nt	nt
T7 RNAP V783M	++	++	nt	nt
PUP	++	nt	nt	nt

- no incorporation, + incorporation to a small extent, ++ efficient incorporation, nt – not tested

Competitive incorporation assays allowed to elucidate the level of discrimination against base-modified terminators characteristic to various polymerases (Fig. 3.1). Structurally different (Das & Georgiadis, 2004) M-MLV-based SuperScript IV and HIV RTs demonstrated different behaviors: HIV RT showed nearly equal rate of incorporation of both dTTP and ddU^{ON2}TP while SuperScript IV displayed clear discrimination against OTDDN. This result is consistent with previously reported substantially more error-prone DNA-dependent cDNA synthesis by HIV RT as compared to M-MLV and AMV RTs, however the authors found that these differences were not so profound on RNA templates (Sebastián-Martín *et al.*, 2018).

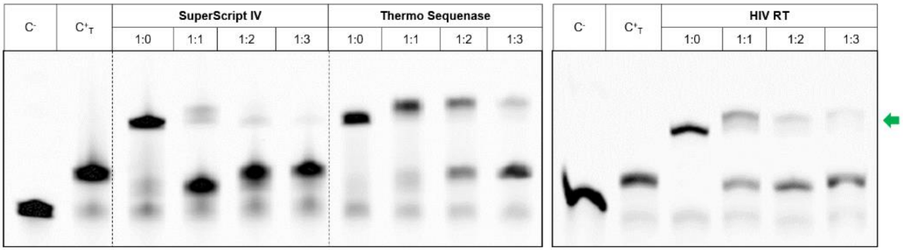


Figure 3.1. Competitive incorporation assay. PEX experiments were performed with Dup^{A} primer-template duplex either with ddU^{ON2}TP alone (1:0) or with various ddU^{ON2}TP ratios relative to dTTP (1:1, 1:2, 1:3). Different enzymes show different levels of discrimination against OTDDN. C^{-} - negative control, C$^{+}_{T}$ – positive control, green arrow indicates OTDDN-labeled products.

Both SuperScript IV and HIV RTs were tested in oligo(dT) primer extension and termination by OTDDNs on mRNA templates with further amplification of labeled cDNA fragments. HIV RT catalyzed the incorporation of OTDDNs more efficiently than SuperScript IV which is evident from the accumulation of short labeling products (Fig. 3.2), however SuperScript IV was selected for further development of transcriptome sequencing techniques. The efficiency of competitive OTDDN incorporation exhibited by SuperScript IV appeared to be sufficient for the generation of cDNA fragment libraries even from small amounts of RNA (see chapter 3.3), moreover, SuperScript IV synthesizes cDNA with higher fidelity than HIV RT which allows to obtain higher quality sequencing data, and some discrimination against OTDDN prevents the extensive formation of labeled fragments with extra-short inserts that do not generate informative sequencing data.

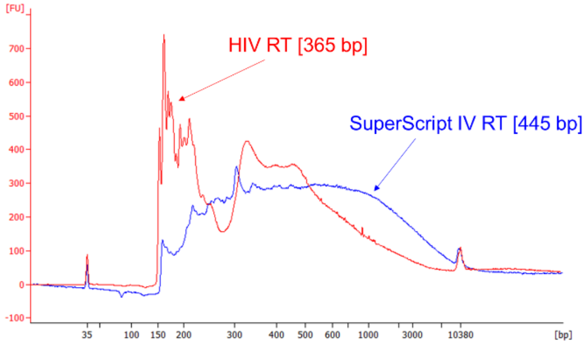


Figure 3.2. The libraries of cDNA fragments produced by SuperScript IV and HIV RTs in the presence of 1:5 ratio of ddU^{ON2}TP and ddC^{ON2}TP to their natural dNTP counterparts. The numbers in brackets stand for the average length of the obtained cDNA fragments measured in the range from 100 bp to 1000 bp.

Interestingly, Thermo Sequenase as well as T7 RNAP mutant preferred OTDDNs over natural dNTPs, with ~2-fold and >1000-fold better usage of tested OTDDNs than unmodified dNTPs, respectively. Thermo Sequenase appeared to be an excellent candidate for DNA fragment library preparation by primer extension and labeling with OTDDNs from dsDNA templates.

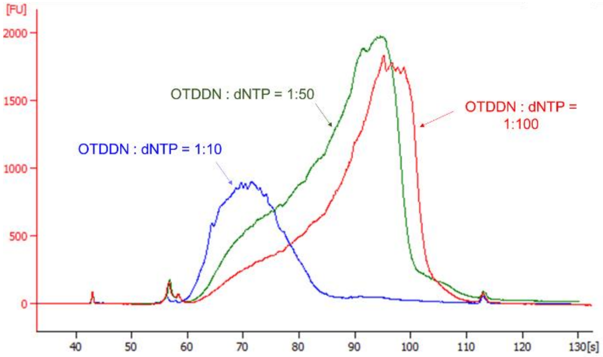


Figure 3.3. The dependency of fragment size distribution on the ratio of OTDDN to its natural dNTP counterpart. The Bioanalyzer traces show three DNA fragment libraries obtained by primer extension with Thermo Sequenase and termination with ddU^{ON2}TP and ddC^{ON2}TP (here, both OTDDNs were used in a single reaction in equimolar amounts). The lower the total amount of OTDDNs is, the longer is the average fragment size.

This notion was tested by primer extension on complex gDNA templates with various ratios of OTDDNs to dNTPs – stochastic nature of OTDDN incorporation should enable the control over the average length of labeled DNA fragments by modulating this ratio. Indeed, the average length of DNA fragments was dependent on the relative amounts of OTDDNs to dNTPs. When higher excess of dNTPs was used, the rates of OTDDN incorporation were lower and thus longer fragments were obtained (Fig. 3.3). The ability to regulate the average insert size makes NGS library preparation method more flexible in terms of compatibility with sequencing instrumentation and with sequencing applications.

3.1.2 Biocompatibility of unnatural triazole linkage

In this study, two configurations of OTDDNs were considered: where dideoxynucleotide and ON were attached (i) base-to-phosphate (Fig. 2.1 B; modifications ON2, ON3, ON4, ON5, ON6, ON8, ON9) or (ii) base-to-base (Fig. 2.1 C; modifications ON1, ON7, ON10). The latter option was eventually rejected because C8-alkyne motif attached to a nucleobase within an ON was only available for pyrimidines which in turn forced us to introduce an additional base into the Illumina adapter sequences. On the other hand, base-to-phosphate linkage was sequence-independent and thus allowed for seamless tagging of sequenceable inserts. This configuration was further tested for biocompatibility, i.e. the ability of various polymerases to synthesize the complementary DNA strand through the unnatural linker using ON modification as a priming site. We identified exonuclease-deficient polymerases able to read through the linker (Table 3.2).

Table 3.2. Read-through activity of various DNA polymerases

Polymerase	Conditions	Read-through
Klenow fragment (exo-)	30ºC 30 min	+
Thermo Sequenase	95ºC 1 min, 60ºC 1 min, 72ºC 15 min	+
Platinum™ SuperFi™		-
Phusion U		-
Phusion (exo-)		++
Phusion (exo-)	95ºC 1 min, 60ºC 1 min, 72ºC 10 min	++
Phusion (exo-)	95ºC 1 min, 60ºC 1 min, 72ºC 5 min	+
SuperScript™ IV	50ºC 30 min	+

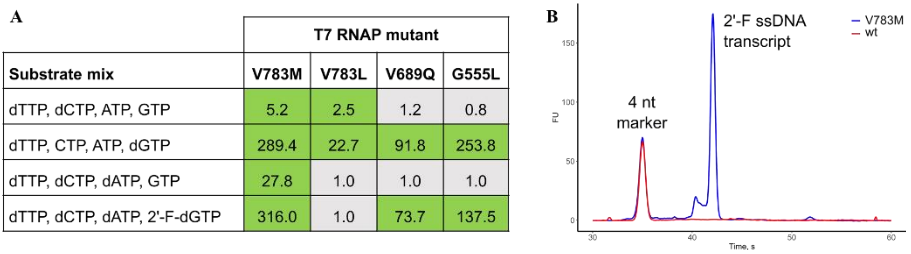
- undetectable read-through, + inefficient but detectable read-through (up to 10% efficiency), ++ good read-through activity (>20% efficiency)

We identified Phusion (exo-) as a best performing enzyme with the single-cycle read-through efficiency of 25-30%. The efficiency depended on the reaction time, Phusion (exo-) reached its maximum single-cycle efficiency in 10 min reaction. Having in mind that most of the OTDDN-based NGS library preparation applications require PCR to introduce full-length adapters, Phusion (exo-) was a fortunate discovery because of its thermal stability and compatibility with other Phusion-based polymerases. This in turn allowed the use of polymerase blends for the amplification of OTDDN-containing libraries (see Methods).

3.2 Labeling of linear amplification products

DNA and RNA polymerases have evolved mechanisms to efficiently discriminate against substrates containing non-cognate sugar moieties. Understanding the determinants of substrate selection would enable the expansion of catalytic properties, of which the transformation of RNA polymerase into DNA polymerase is of great interest due to the possibility to isothermally synthesize ssDNA.

To create a new method for linear DNA amplification with an integrated opportunity to label amplification products for subsequent high-throughput sequencing, *in vitro* evolution was used to select T7 RNAP variants able to catalyze the incorporation of dNTPs and thus synthesize highly stable transcripts. Mutant variant V783M exhibited the best performance with substrate mixture consisting of dTTP, dCTP, dATP and 2′-F-dGTP (Fig. 3.4) and was used for further feasibility experiments. Neither of tested mutants were able to synthesize unmodified ssDNA products most probably because of intrinsic requirement for GTP to stabilize the initiation complex (Koh *et al*., 2018). 2′-fluoro modification does not substantially change transcript conformation as compared to RNA because highly electronegative fluorine atom confers an RNA-like C3′-endo conformation of nucleotide sugar moiety (Manoharan, 1999; Zhu *et al*., 2015), thus substitution of GTP to 2′-F-dGTP appeared to be highly successful.



A

	T7 RNAP mutant			
Substrate mix	V783M	V783L	V689Q	G555L
dTTP, dCTP, ATP, GTP	5.2	2.5	1.2	0.8
dTTP, CTP, ATP, dGTP	289.4	22.7	91.8	253.8
dTTP, dCTP, dATP, GTP	27.8	1.0	1.0	1.0
dTTP, dCTP, dATP, 2'-F-dGTP	316.0	1.0	73.7	137.5

Figure 3.4. Selection of T7 RNAP variants with reduced substrate discrimination. **(A)** – T7 RNAP mutants able to incorporate various combination of NTPs and dNTPs as well as to synthesize chimeric ssDNA containing 2′-F-dG. **(B)** – a capillary electrophoresis trace of a transcript consisting of dTTP, dCTP, dATP, and 2′-F-dGTP synthesized by V783M mutant.

Isothermal amplification of a portion of intact plasmid DNA downstream of T7 promoter in the presence of dNTPs and OTDDNs enabled to produce libraries of randomly terminated labeled transcription products. Subsequent PCR amplification using IVT products as templates successfully generated sequencing-ready samples. Interestingly, empirical screening for optimal OTDDN to dNTP ratio revealed high preference of T7 RNAP mutant to OTDDNs – libraries of average size of ~300 bp were obtained with 5000× deficiency of OTDDN relative to respective dNTP. On average, 93.5% (89.7-95.2% among replicates) of reads aligned to the reference plasmid sequence and covered an expected region immediately downstream of T7 promoter (Fig. 3.5). Read structure was of the expected composition: first 8 bases of forward reads had random base distribution attributable to the randomized region within OTDDN, while the dominance of G and A bases at the 9th position corresponded to ddCTP and ddUTP incorporation sites, respectively (Fig. 3.6).

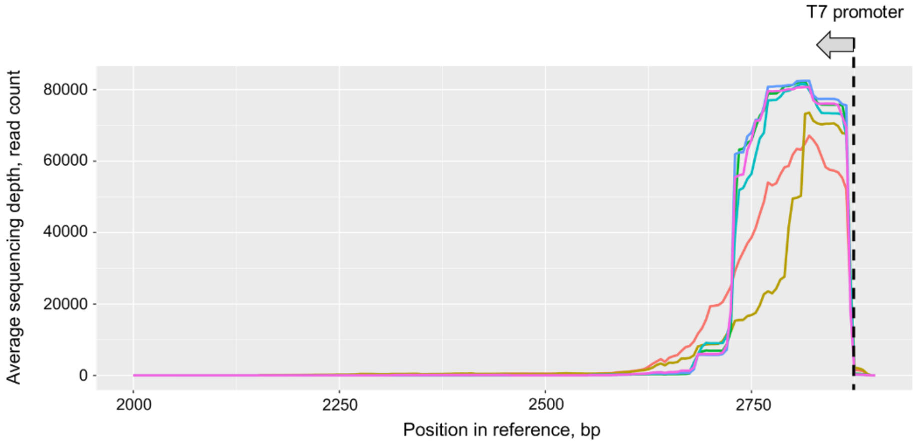


Figure 3.5. Sequencing of directly PCR-amplified IVT products synthesized by T7 RNAP V783M using dTTP, dCTP, dATP, 2′-F-dGTP and OTDDN. The obtained read coverage of pTZ19R template plasmid. Each line graph represents an individually prepared sample.

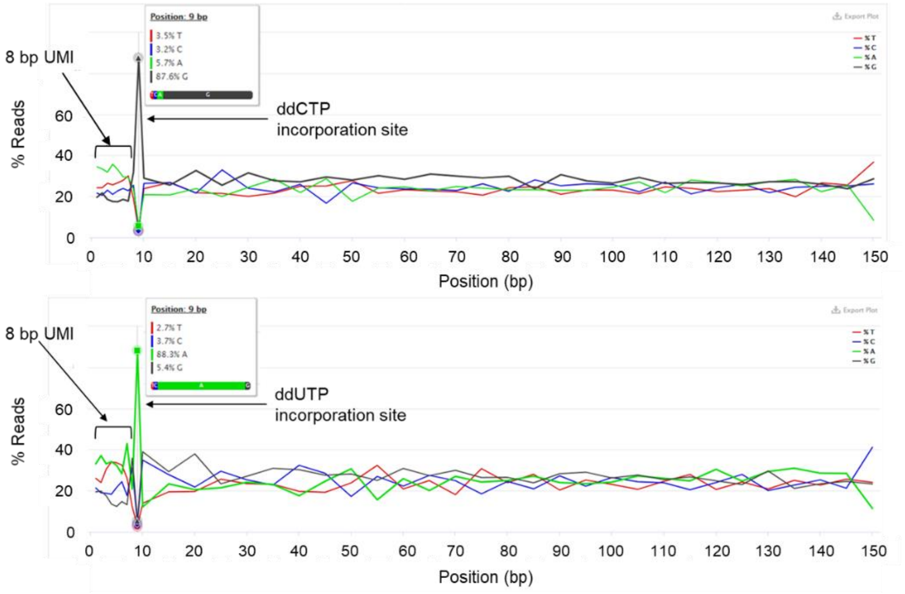


Figure 3.6. Base composition of forward reads obtained upon sequening of OTDDN labeled 2′-F ssDNA transcripts. First 8 bp correspond to the randomized region within OTDDN, and the base at 9th position should be complementary to the dideoxynucleotide. Clear dominance of G and A bases at 9th position confirm the incorporation of ddCTP and ddUTP, respectively.

The proposed approach opens doors for convenient linear isothermal amplification of DNA with introduced T7 promoter sequences and simple downstream processing and handling of highly stable transcripts in 2′-F ssDNA form.

3.3 OTDDNs for DNA sequencing applications

DNA sequencing can be global, when all DNA molecules in a sample are sequenced, e.g. in whole genome or metagenome sequencing applications, or targeted, when sample is enriched for loci of interest, e.g. by sequencing specific gene panels. Both approaches are compatible with OTDDN technology. For whole genome sequencing, OTDDNs offer a convenient protocol that accepts intact DNA as an input. Typically, DNA fragmentation prior to adapter addition is performed either by physical methods, such as acoustic shearing, or enzymatic methods, such as digestion by non-specific endonucleases (Head *et al*., 2014). Integration of fragment generation together with adapter addition step eliminates the need for specialized equipment or nucleic acids cleaving enzymes. As for targeted approach, the stochastic OTDDN incorporation enables the unique opportunity to terminate the extension of specific primers in the *a priori* unknown genomic regions nearby. This feature has a number of advantages over traditional PCR enrichment as will be exemplified by 16S rRNA gene sequencing for microbiome characterization.

3.3.1 Labeling of primer extension products

Whole genome sequencing by extension and OTDDN labeling of random primers was tested on *Escherichia coli* gDNA. Intact DNA was denatured allowing the annealing of random primers which were then extended and tagged by OTDDN. This step generated DNA fragments (average size ~350-550 bp depending on the concentration of random primers and OTDDNs) suitable for sequencing, thus eliminating the need for a separate fragmentation step. Amplified libraries were subjected to Illumina paired-end sequencing.

85% of reads aligned to the reference genome and covered the whole *E. coli* chromosome (Fig. 3.7 A). The GC coverage was in good agreement with the GC content reported for the *E. coli* genome (Fig. 3.7 B). Random priming enables an additional layer of control over the insert size through the modulation of random primers to gDNA ratio. Here, keeping OTDDN to dNTPs ratio constant, we tested two concentrations of random primers and indeed observed that higher concentration of primers resulted in denser annealing pattern and thus shorter inserts (Fig. 3.7 C).

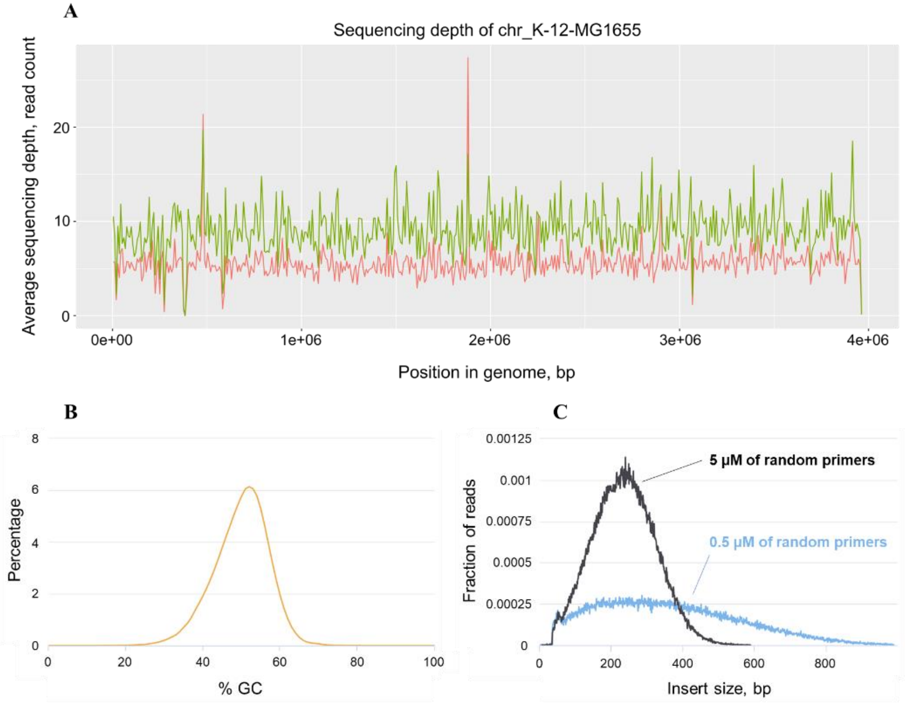


Figure 3.7. *Escherichia coli* genome sequencing by random primer extension and termination with OTDDNs. **(A)** – *E. coli* genome coverage obtained in proof-of-principle experiment. Green and red lines represent two different libraries. **(B)** – the obtained sequence GC content distribution agrees well with that reported for the *E. coli* genome (GC = 50.79%; UCSC Genome Browser, Kent *et al*., 2002). **(C)** – insert size distribution depends on the ratio of random primers to the template. The graph shows obtained insert sizes in libraries prepared from 100 ng of *E. coli* genomic DNA and different amounts of random primers.

Labeling of specific primer extension products was first tested on a low-complexity M13mp18 viral genome. Two specific primers oriented in the same direction and targeting loci 2980 bp apart from each other were both annealed and extended in a single reaction. OTDDN-labeled extension products were sequenced, with priming sites comprising the beginning of forward reads. 71% of reads aligned to the reference genome, and the obtained median insert size was 191 bp. The coverage profile matched the theoretical design very well: 5′-terminal regions of sequenced inserts aligned at fixed positions that corresponded to target loci, while 3′-terminal regions were dispersed indicating random OTDDN incorporation events (Fig. 3.8 A). We termed this approach *semi-targeted sequencing*, emphasizing that only a single target-specific primer is used to generate sequencing-ready amplicons.

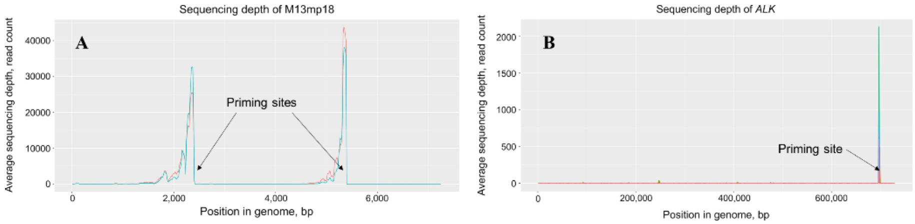


Figure 3.8. Semi-targeted sequencing with OTDDNs. **(A)** – the coverage of M13mp18 genome concentrates in two loci that correspond to two target sites. The fixed termini of sequenced inserts correspond to the priming sites while randomly distributed termini illustrate the stochastic nature of OTDDN incorporation. Each line graph represents a technical replicate. **(B)** – the coverage of human *ALK* gene which was obtained upon sequencing of semi-targeted libraries. Each line graph depicts results obtained with each of the 5 tested *ALK*-specific primers. The length of *ALK* is ~730 kb, thus the coverage graphs appear as a peak.

Single primer extension-based approaches are more prone to specificity issues because each mispriming event will be captured in the final library. Semi-targeted sequencing was challenged with complex human gDNA template. Five primers specific to *ALK* gene were designed and tested individually. Surprisingly, one of the primers (ALK-2, see Table 2.6) generated ~73% on-target reads (68-80% among different samples) while others were not able to reach 50%. As could be expected, on-target reads concentrated near the priming locus (Fig. 3.8 B). Considering the size of the human genome, obtaining ~73% of correct reads from a single primer extension is remarkable, and encourages the application of semi-targeted approach for complex templates.

3.3.2 Semi-targeted sequencing of 16S rRNA gene

16S rRNA gene is widely used to differentiate operational taxonomic units (OTUs) for the profiling of microbial communities. High-throughput sequencing of 16S rRNA amplicons led to rapid growth of available gene sequence data, which to this day outnumbers complete genome assemblies. Despite its versatility, intragenomic heterogeneity of 16S rRNA gene copies impairs classification accuracy as well as quantitative representation of microbial communities. We developed a new semi-targeted 16S rRNA gene sequencing method (st16S-seq) which directly links each 16S rRNA gene copy with adjacent genomic locus upstream of the gene and enables highly accurate classification and unambiguous quantification of taxa (Kapustina *et al*., 2021a).

Diagnostic potential of the proposed approach was first evaluated *in silico* by the analysis of publicly available 13 570 unique bacterial genome assemblies. Indeed, the inclusion of near-16S regions increased the classification accuracy at species level when strains were left out (Fig. 3.9). The accuracy increased with increasing length of included genomic fragment up to 400 bp. The mean classification accuracy using 16S rRNA sequences alone was comparable to that of near-16S regions linked to V1-V2. For individual hypervariable regions, the classification accuracy was substantially lower as compared to both full-length gene and near-16S linked to V1-V2.

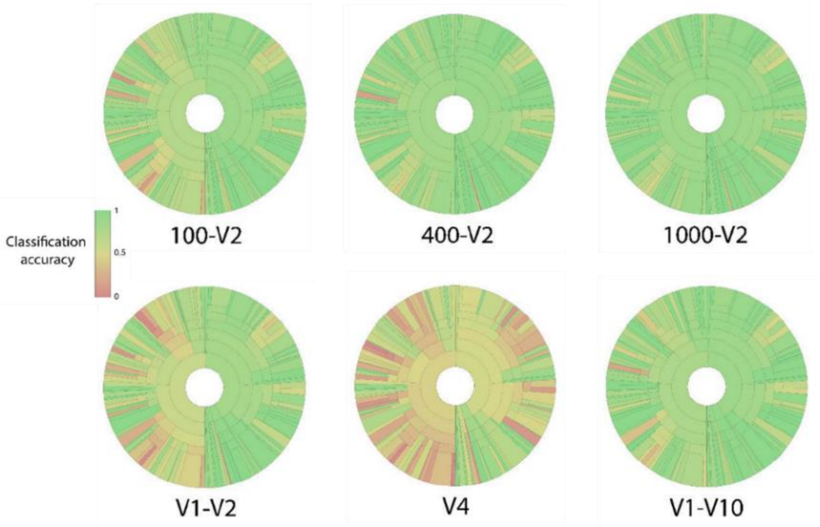


Figure 3.9. Krona charts depicting *in silico* estimated classification accuracy at the species level. The outer ring corresponds to the genus/family level. The size of circular fragments is proportional to the number of sequences belonging to the rank. For sequences upstream of 16S rRNA gene the length of included genomic fragment is indicated (100 bp, 400 bp, 1000 bp). In all cases near-16S regions were linked to V1-V2 16S rRNA sequences.

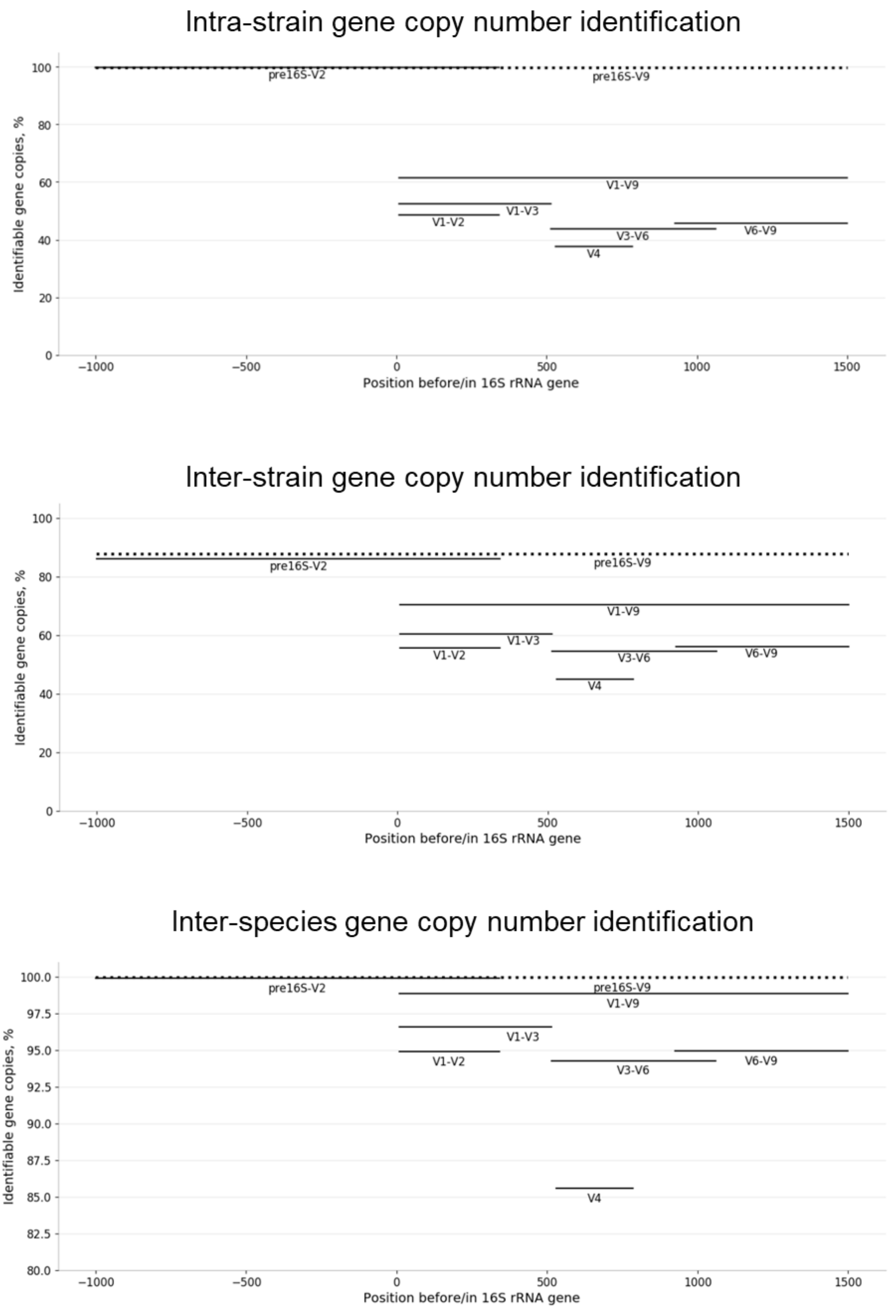


Figure 3.10. The percentage of identifiable 16S rRNA gene copy numbers as assessed by various regions of 16S rRNA gene. V1-V9 indicate 16S rRNA hypervariable regions that were included in the analysis.

The ability to identify 16S rRNA gene copy numbers would significantly improve quantification of taxa. Within individual genomes, 16S rRNA hypervariable regions were able to differentiate between up to ~50% of gene copies. The full-length gene sequence raised the fraction of identifiable copies to 60%. The inclusion of near-16S region in conjunction with either V1-V2 or V1-V9 increased the discrimination to 99.7%. Microbiome samples often contain a mixture of closely related bacterial lineages which complicates quantitative assessment because of the presence of identical 16S rRNA genes in genomes of different strains. To evaluate inter-strain 16S rRNA gene copy differentiation, we included species with at least two strains in the analysis. Although neither sequence allowed for absolute discrimination, we observed >80% of identifiable 16S rRNA copies for near-16S and 16S rRNA sequence combination in contrast to ≤70% for within-gene sequences. At higher taxonomic level, inter-species sequence variation somewhat lowers the discriminatory power of full-length 16S rRNA sequence and sequences of hypervariable regions, however there is almost no impact on near-16S region thus allowing for absolute gene copy number identification (Fig. 3.10).

Practical st16S-seq feasibility studies were first performed on well characterized *E. coli* genome which is known to contain 7 copies of rRNA operon (Maeda *et al*., 2015). st16S-seq analysis was able to identify all 7 copies by the assembly of 16S rRNA proximal regions. Likewise, almost all gene copies were identified within the genomes of 8 prokaryotic members of ZymoBIOMICS™ mock microbial community (Fig. 3.11). In some instances, not all gene copies were resolved because intragenomic sequence differences occur at a marginal distance which can be captured and reliably sequenced by st16S-seq using short reads. In bacteria with multiple 16S rRNA gene copies, the evolution of 16S rRNA genes is thought to occur not only by vertical transmission of mutations, but also by non-reciprocal recombination with either horizontally acquired or intragenomic donors (Hashimoto *et al.*, 2003; Espejo, Plaza, 2018). Intragenomic recombination events might in turn result in duplications of the chromosomal regions nearby bringing more complexity to the identification of gene copy numbers.

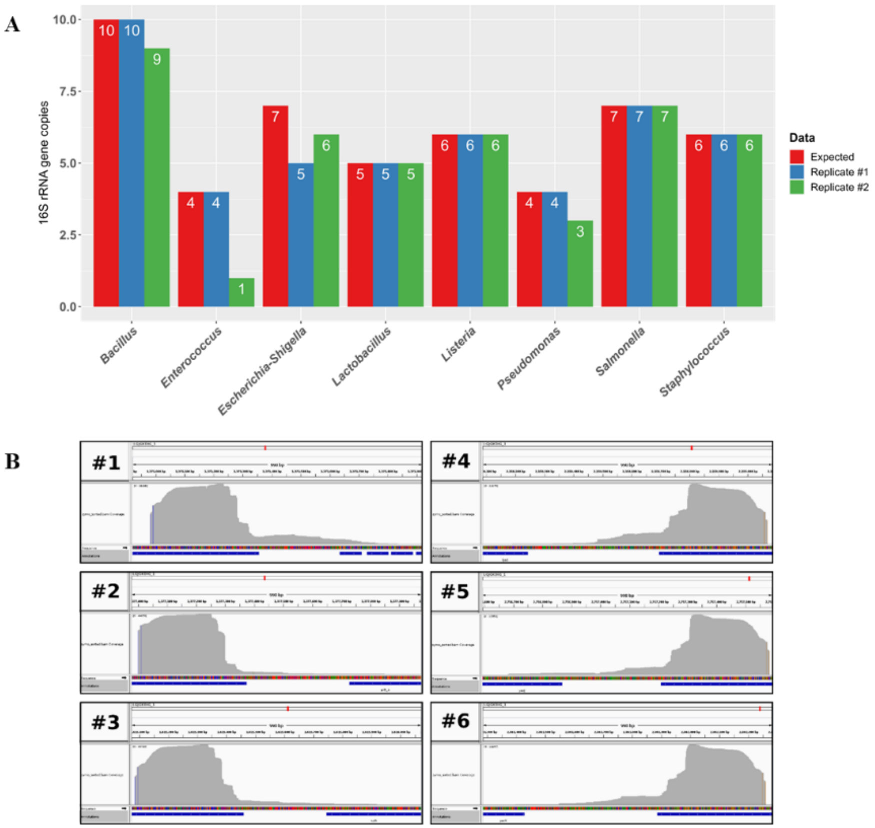


Figure 3.11. 16S rRNA gene copy number identification from st16S-seq data. **(A)** – the number of 16S rRNA gene copies detected within genomes of the members of ZymoBIOMICS™ Microbial Community DNA standard that equates to the number of 16S rRNA contigs after removal of artifactual sequences. **(B)** – read coverage of each of the six 16S rRNA gene copies and upstream regions within *Listeria monocytogenes* genome.

Identification of 16S rRNA gene copy numbers allows to normalize read counts for accurate quantitative estimation of taxa. st16S-seq data strongly correlated with the expected abundance distribution in two analyzed mock communities, with Pearson′s correlation coefficients of 0.96-0.97 and 0.88-0.90 for ZymoBIOMICS (Fig. 3.12 A) and ATCC (Fig. 3.12 B) standards, respectively. In contrast, PCR-based techniques demonstrated poorer (Quick-16S and NEXTFLEX V1-V3) or inconsistent (EMP, NEXTFLEX V4, QIAseq and Swift) performance as compared to st16S-seq, which is attributable to widely acknowledged limitations posed by PCR primer design and unequal discriminatory power of individual 16S rRNA variable regions (Klindworth *et al*., 2013; Winand *et al*., 2019; Soriano-Lerma *et al*., 2020).

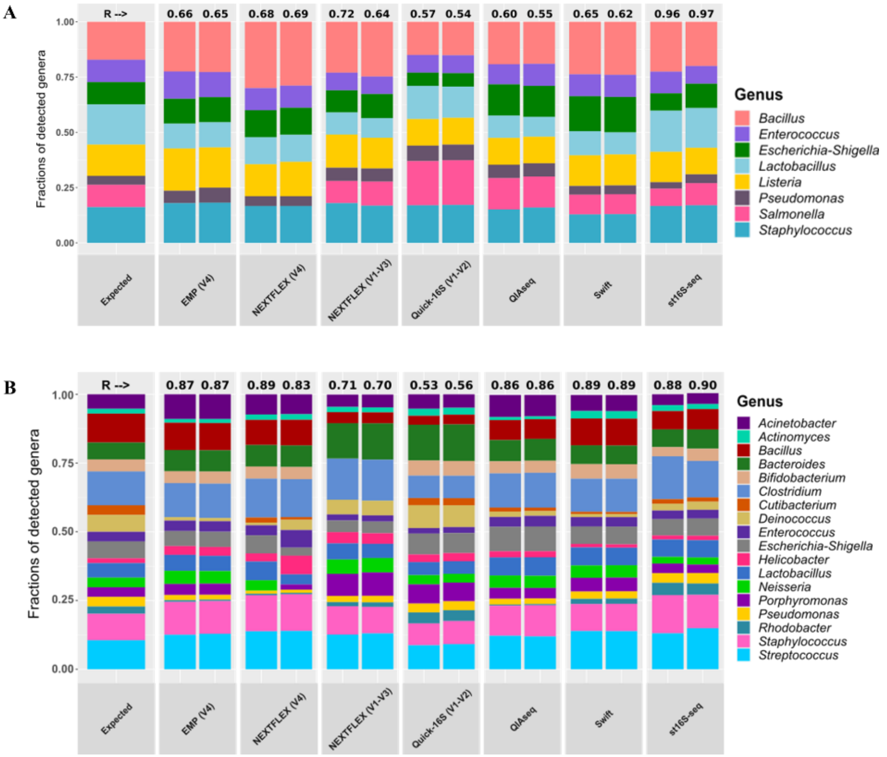


Figure 3.12. Validation of st16S-seq on mock community DNA standards and comparison with conventional techniques. **(A)** – read distribution across bacterial genera in libraries prepared from ZymoBIOMICS™ Microbial Community DNA standard with various commercially available kits and st16S-seq approach. The numbers above the bars indicate Pearson′s correlation coefficients between the expected and obtained read distributions. Two replicates are shown for each sample. **(B)** – read distribution across bacterial genera in libraries prepared from ATCC microbiome standard (ATCC MSA-1002™) DNA with various commercially available kits and st16S-seq approach. The numbers above the bars indicate Pearson′s correlation coefficients between the expected and obtained read distributions. Two replicates are shown for each sample.

The inclusion of sequences upstream of 16S rRNA gene into sequencing library improved the classification accuracy at species level. We have assessed the ability of targeted methods and st16S-seq to discern bacterial species within mock communities employing both unmerged and merged paired-end reads for the analysis. The use of unmerged reads placed st16S-seq on a par with V1-V2-containing amplicons. In contrast, the analysis of assembled genome-linked contigs in st16S-seq datasets allowed to correctly identify all members of mock communities, except for *Streptococcus agalactiae*, while the precision of PCR-based methods did not improve from the use of merged reads.

To analyze the utility of st16S-seq for the characterization of highly complex communities, we sequenced libraries prepared from soil-derived DNA. Principal component analysis (PCA) considering the relative abundances of reads assigned per bacterial species across soil samples sequenced using different methods revealed that data based on V4 and NEXTFLEX V1-V3 amplicon sequencing form distinct clusters, while st16S-seq and Quick-16S can approximate the variability of read fractions detected by WGS. Moreover, when st16S-seq data is analyzed as pseudocontigs, meaning that all reads associated with individual OTUs are analyzed as a whole, st16S-seq clusters with WGS even better (Fig. 3.13).

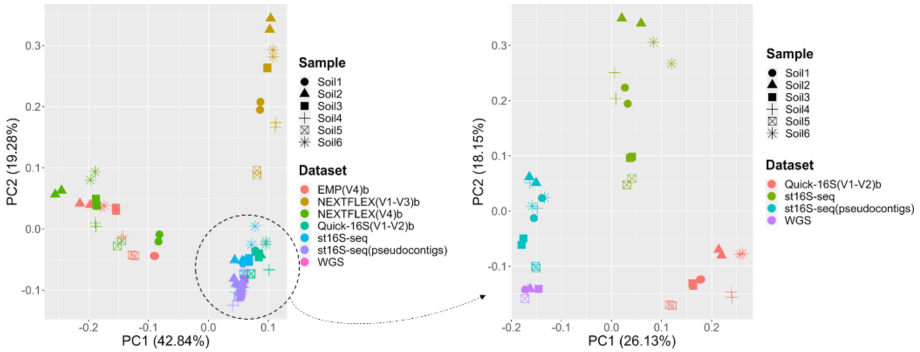


Figure 3.13. Species-level discriminatory power of st16S-seq on complex samples and comparison with conventional techniques. PCA analysis considering the relative abundance of reads assigned per bacterial species across different soil samples processed by various library preparation techniques. The graph on the right depicts data that clusters near WGS.

Taken together, we demonstrated that direct linking of adjacent genomic sequences to those of V1-V2 16S rRNA region gives a number of significant methodological advancements to microbiome characterization.

3.4 OTDDNs for RNA sequencing applications

While direct RNA sequencing is possible (Ozsolak *et al*., 2009; Garalde *et al.*, 2018), currently most RNA-seq experiments are performed on instruments that sequence DNA molecules due to their technical maturity. This means that to prepare RNA-seq library RNA must be converted into cDNA of certain size flanked by adapter sequences. The cDNA library preparation methods may vary depending on the RNA species under investigation, which in turn can differ in size, sequence, structural features and abundance (Hrdlickova *et al*., 2017). Here, we demonstrate that OTDDNs can be applied for rapid, simple and accurate cDNA library preparation and ensure the retention of strand specificity, which is important for the identification of antisense expression and novel RNA species. Our efforts focused on the most common application of RNA-seq – sequencing of polyadenylated RNA, either throughout the whole transcript length or targeting terminal regions.

3.4.1 Labeling of cDNA

To test the utility of OTDDN-based cDNA labeling approach for sequencing mRNAs along their full length, oligo(dT)-primed first strand cDNA synthesis was performed with well-characterized UHRR total RNA. Then, second strand synthesis was primed by random oligonucleotides and extension products were labeled by OTDDNs.

On average, 95.3% (94.4-97.2% among replicates) of sequencing reads mapped to the human genome. With 0.9 M reads we detected over 17 100 unique genes, with the mean strand specificity value of 98.4% (97.8-98.7% among replicates). The read coverage of transcripts was slightly biased towards 3′ terminus which is attributable to the oligo(dT)-primed RT step. Except for this, the gene body coverage was comparable to that obtainable by the conventional technique (Fig. 3.14).

The analysis of detected transcript biotypes exposed an area for further improvement – substantial contamination with rRNA reads (on average 42.8%) was observed. This might be explained by the capture of intragenic polyA stretches within rRNA transcripts during the RT step. Specificity for mRNAs could be improved by optimizing RT conditions or by performing mRNA enrichment from total RNA before the library preparation. Alternatively, not-so-random primers (Armour *et al*., 2009) might be used to prepare tagged cDNA fragments during RT with simultaneous depletion of abundant RNA species.

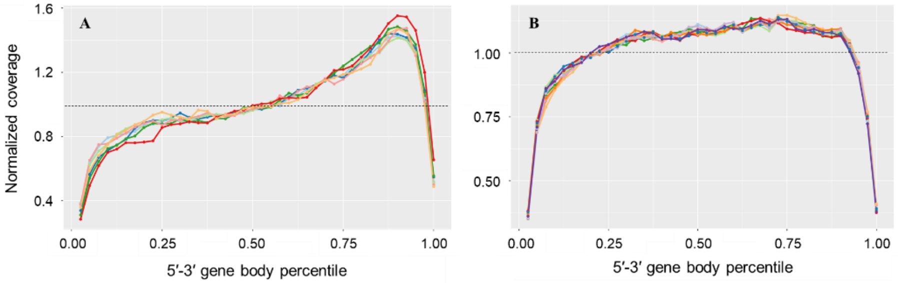


Figure 3.14. RNA sequencing along the whole transcript length. **(A)** – gene body coverage obtained in libraries prepared with OTDDNs. **(B)** – gene body coverage typical to standard RNA sequencing. Here, libraries were prepared with the Collibri™ Stranded RNA Library Prep Kit. Each line represents a technical replicate. The ideal coverage value of 1 is shown with a dashed line.

One interesting opportunity raised by OTDDN labeling is the preparation of cDNA libraries without PCR amplification. To do so, RT primer must have a full-length adapter sequence at the 5′ terminus and OTDDN must contain a full-length adapter modification. Moreover, bridge amplification on the surface of a flow cell must be compatible with templates containing unnatural linkages. We were able to synthesize OTDDN having a 67 nt oligonucleotide modification (ddC^{ON10}TP) which corresponded to an indexed P7 Illumina adapter. Model experiments (see chapter 2.4.1) revealed that Illumina random clustering chemistry tolerates triazole-based linkage within DNA template very well: within an equimolar sample pool, OTDDN-containing sample constituted 42.0% of all identified reads, while control samples constituted 40.9% and 17.1%. If polymerase used for bridge amplification would not be able to synthesize DNA through the OTDDN linker, we would expect to retrieve disproportionately less clusters for OTDDN sample than for controls.

Next, we tested PCR-free library preparation from UHRR by oligo(dT) primer extension and labeling. As could be expected, 93% of obtained reads corresponded to human protein coding genes, and the majority of reads mapped near the 3′ termini. We observed that all reverse reads started with A and G bases that correspond to AldU and ddCTP, respectively. Further exploration of this PCR-free opportunity would enable easiest possible sequencing-ready RNA-seq library preparation.

3.4.2 Gene expression profiling

Whole transcriptome sequencing generates the most comprehensive transcriptomic datasets, however the sensitivity and accuracy of detection of relative changes in gene expression across sample groups is hindered by read coverage bias towards longer transcripts (Mortazavi *et al*., 2008; Gao *et al*., 2011). While long-read sequencing technologies allowing full-length transcript analysis, such as Iso-Seq, may solve this issue by producing single read per transcript with no tradeoff in regard to structural information, currently this approach is mostly adopted to study non-model organisms (Minio *et al*., 2019; Chen *et al*., 2019). For gene expression profiling employing short-read sequencers, library preparation techniques that generate only one fragment per transcript, usually adjacent to either 5′ or 3′ terminus, are acknowledged as a good cost-effective alternative to whole transcriptome RNA-seq and were rapidly adopted for high-throughput single-cell sequencing (Xiong *et al*., 2017; Ma *et al*., 2019; Islam *et al*., 2012).

We created and tested the workflow for the capture of mRNA 5′ termini, which is based on reverse transcription with template switch and subsequent second strand synthesis in the presence of OTDDNs. Second strand synthesis was primed targeting the template switch oligonucleotide which labeled the 5′ terminus of the original mRNA. 83% of sequenced reads mapped to human genome, and among those 78% of reads corresponded to protein coding genes. Proof-of-principle experiment detected 10 233 unique genes with strand specificity of 91.1%. Gene body coverage analysis revealed the increased coverage of 5′ termini, however 3′ termini were also overrepresented (Fig. 3.15).

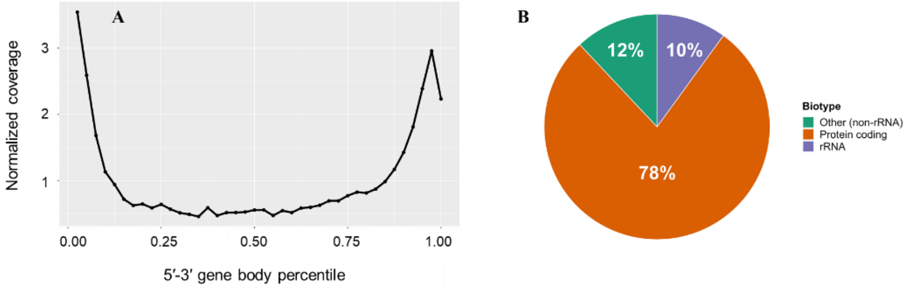


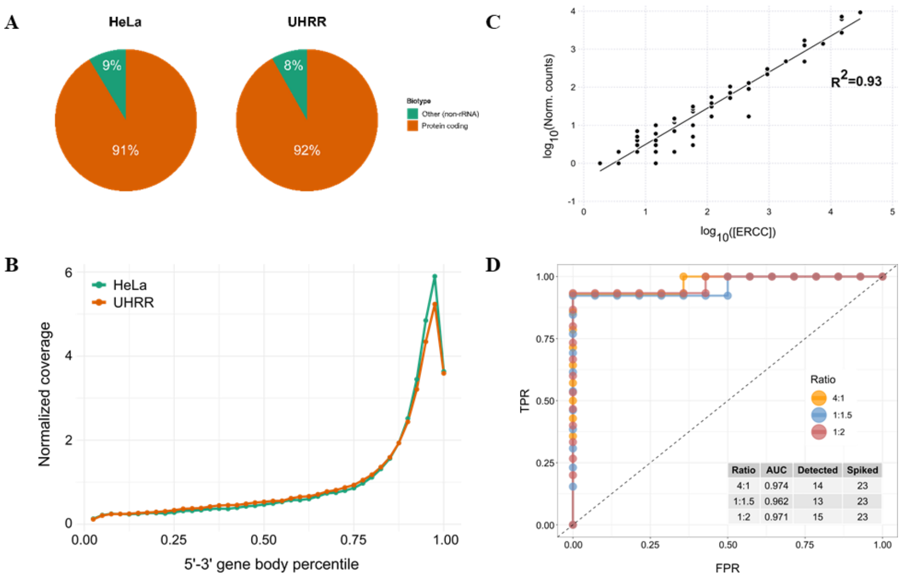
Figure 3.15. 5′-end RNA sequencing with OTDDNs. **(A)** – the obtained gene body coverage. 5′ terminal regions are highly represented, however overrepresentation of 3′ termini is also visible. **(B)** – captured transcript biotypes. Most reads, as expected, correspond to protein coding genes.

This peculiar effect might have been caused by strand invasion (Tang *et al*., 2013), i.e. interruption of cDNA synthesis by premature annealing of TSO to cDNA, which results in oligo(dT)-primed cDNA synthesis termination closer to the 3′ end.

The enrichment for 3′ UTRs allows insights about 3′UTR isoform choice variability which was reported to contribute to phenotypic diversity across individual cells (Velten *et al*., 2015). We developed a new method for high-throughput gene expression profiling which generates fragment libraries from the 3′-terminal transcript regions with rapid and simple single-tube protocol. We termed this approach mRNA sequencing via terminator-assisted synthesis, or MTAS-seq.

To validate the technique, we sequenced MTAS-seq libraries prepared from well characterized HeLa and UHRR total RNA samples spiked with ERCC transcript mixes, with three technical replicates per RNA input, which ranged from 0.5 ng to 500 ng. The obtained libraries were of similar size indicating that OTDDN incorporation rate is robust across different RNA inputs given the same OTDDN ratio to corresponding dNTPs (Supplementary Fig. 2 A-B). 99.4 – 99.8% of sequencing reads from each sample mapped to the human genome and ERCCs after UMI trimming, with strand specificity of >99 %. We obtained sequences for more than 19 000 genes in UHRR samples and nearly 15 000 genes in HeLa samples with only 2 M reads. Reverse transcription conditions demonstrated high specificity for mRNAs even though starting material was total RNA: there were virtually no traces of rRNA reads indicating no mispriming events, and read coverage, as expected, concentrated at the 3′ terminal region of RNA transcripts (Fig. 3.16 A-B).

To assess dose-response in MTAS-seq libraries, we compared the detected ERCC counts to expected ones and observed that with at least 50 unique ERCC transcripts identified the correlation (R^2) with the expected distribution was 0.91-0.94 (Fig. 3.16 C). We next evaluated the discriminatory power of differential expression detection by assessing ERCC ratio performance with receiver operating characteristic (ROC) curves and area under the curve (AUC) statistics. With at least 13 ERCC spikes detected per abundance ratio, AUC analysis indicated good diagnostic power of MTAS-seq assay, with AUC values >0.96 for all ratios (Fig. 3.16 D). This suggests the utility of MTAS-seq for highly accurate gene expression profiling, with an additional advantage of UMI labeling which is especially important for low-input applications prone to high PCR duplication rates (Supplementary Fig. 2 C-D).



Ratio	AUC	Detected	Spiked
4:1	0.974	14	23
1:1.5	0.962	13	23
1:2	0.971	15	23

Figure 3.16. Data quality obtained with MTAS-seq. **(A)** – RNA species captured in MTAS-seq libraries prepared from well characterized RNA (note that “Other” category includes ERCC RNA Spike-Ins which were captured via their polyA tails). **(B)** – typical gene body coverage. **(C)** – the correlation coefficient (R^2) of detected ERCC counts versus expected in MTAS-seq library prepared from 500 ng of UHRR with ~2% of ERCC mix was 0.93, with 55 different ERCCs detected. **(D)** – ROC curves indicate *erccdashboard* analysis to assess the performance of differential expression estimation. FPR – false positive rate, TPR – true positive rate.

To assess whether high-quality libraries might be produced directly from eukaryotic cell lysates leaving behind RNA extraction, we first purified total RNA from HEK-293 cells and determined the approximate amount of RNA per cell, which was ~12 pg. Next, we prepared MTAS-seq libraries from various amounts of HEK-293 cells and, in parallel, from purified RNA which amount corresponded to the cellular RNA contents used in crude lysate experiment. Libraries for each input were prepared in quadruplicates.

On average, 99.1% (98.6-99.3%) of sequencing reads aligned to human genome in crude lysate samples, and 99.2% (96.7-99.5%) - in RNA samples. We observed a good agreement of gene detection capacity between corresponding lysate and RNA samples (Fig. 3.17) as well as good technical reproducibility of library prep from whole cells.

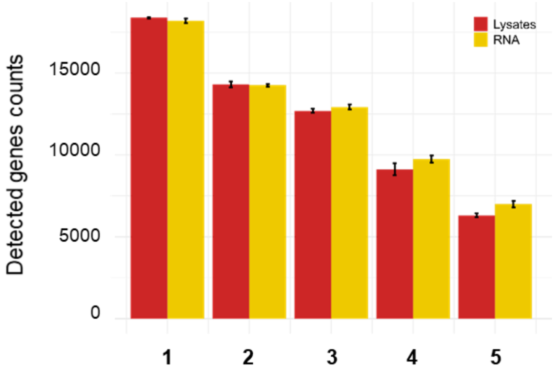


Figure 3.17. Numbers of detected genes in MTAS-seq libraries prepared from different amounts of RNA and cells. 1 – 120 ng RNA or 10 000 cells, 2 – 12 ng RNA or 1000 cells, 3 – 6 ng RNA or 500 cells, 4 – 1.2 ng RNA or 100 cells, 5 – 0.12 ng RNA or 10 cells.

We further applied the direct approach for a different cell type – mouse BALB/3T3 fibroblasts – and obtained high-quality data confirming the reliability and robustness of MTAS-seq as well as the ability to generate libraries from sub-nanogram quantities of total RNA (Fig. 3.18).

The performance of MTAS-seq was compared to that of a conventional commercially available library preparation kit. MTAS-seq clearly outperformed the conventional method in terms of specificity to mRNAs, however detected less genes from very low amounts of starting material (Table 3.3).

Table 3.3. The comparison of technical parameters obtained in conventional 3′ mRNA sequencing dataset (Kit) and OTDDN-based MTAS-seq (OTDDN). Samples were prepared in triplicates; sequencing depth was normalized to 2 M reads/sample. Values are reported as mean ± standard deviation.

Parameter	**UHRR**		**HeLa**	
	OTDDN	**Kit**	**OTDDN**	**Kit**
Mapped reads, %	99.45±0.03	99.05±0.39	99.68±0.05	99.01±0.77
Strand specificity, %	99.21±0.06	98.58±0.30	99.50±0.04	99.14±0.10
Protein coding genes, %	91.71±0.31	74.07±4.74	91.38±0.18	73.90±2.16
rRNA contamination, %	0.06±0.02	7.51±3.94	0.03±0.01	9.51±3.73
Detected genes (500 ng)	19046±23	19124±54	14874±20	15032±10
Detected genes (100 ng)	18745±38	18656±22	14443±22	14794±59
Detected genes (10 ng)	15633±163	15134±926	11877±15	11554±334
Detected genes (0.5 ng)	7273±73	9938±393	5055±55	7388±602

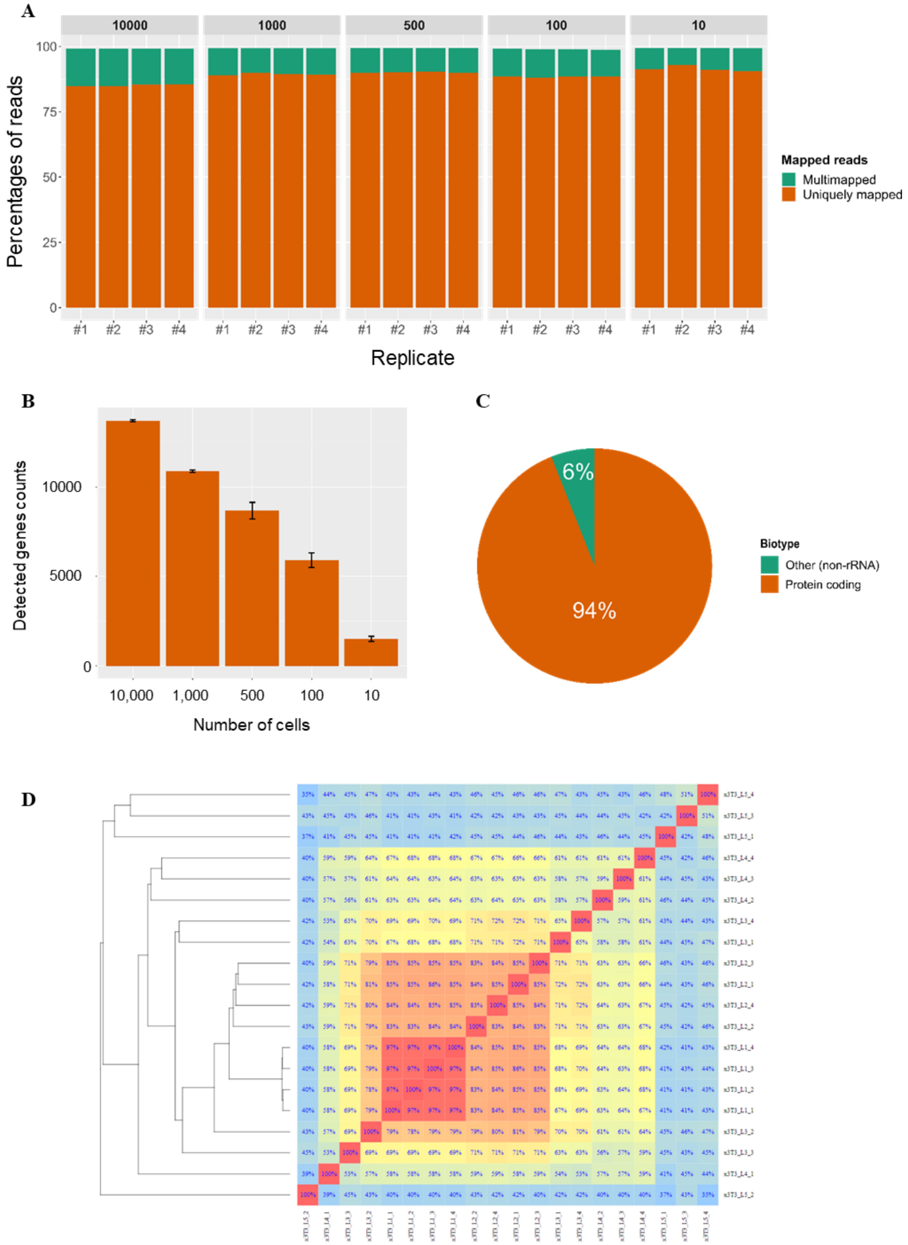


Figure 3.18. Gene expression profiling in BALB/3T3 cell lysates by MTAS-seq. **(A)** – the percentages of reads aligned to mouse genome in libraries prepared from various amounts of cells. **(B)** – the numbers of detected genes in libraries prepared from different amounts of starting material. **(C)** – captured RNA species. **(D)** – gene counts correlation matrix. The percentages indicate Pearson correlation coefficients. L1 samples correspond to 10 000 cells, L2 – 1000 cells, L3 – 500 cells, L4 – 100 cells, L5 – 10 cells.

3.4.3 Single-cell sequencing applications

We reasoned that generation of oligonucleotide-tagged cDNA fragments can be applicable for high-throughput single-cell sequencing, eliminating the need for transcriptome preamplification and subsequent tagmentation. To test this notion, we performed classical species mixing model experiments, using a mixture of HEK-293 and BALB/3T3 cells for encapsulation together with barcoded polystyrene beads originally designed for Drop-seq. The protocol was adapted such as to accommodate optimal reverse transcription conditions for OTDDN incorporation. Indeed, time- and labor-saving advantages of such modified protocol were apparent, with workflow time reduced by ~50% as amplification step readily generated sequenceable cDNA fragment library.

Upon sequencing, 80-90% of reads aligned to combined human and mouse genome reference after quality trimming. Among them, >90 % of identified transcripts were of protein coding genes. We obtained species cross-contamination levels of 2-8 % which are typical for the type of cell processing approach used for proof of principle studies, indicating that modified workflow as such does not introduce any additional undesirable molecular recombination events (Fig. 3.19). The side by side comparison of technical parameters between conventional and OTDDN-based Drop-seq datasets is given in Table 3.4. Similarly to bulk sequencing, here OTDDN-based method underperformed only in terms of gene detection sensitivity. Careful further optimization of reverse transcription and amplification conditions, enabling more efficient incorporation and copying through OTDDN linker would encourage wide adoption of the proposed technique.

Table 3.4. The comparison of technical parameters obtained in standard and modified Drop-seq datasets

Parameter	Standard	Modified
Mapped reads, %	84.4	83.8
Cross-contamination level, %	2.5	2.7
Assigned features, %	55.4	63.0
Transcript capture efficiency, %	11.8	0.5
Protein coding genes, %	93.3	93.5

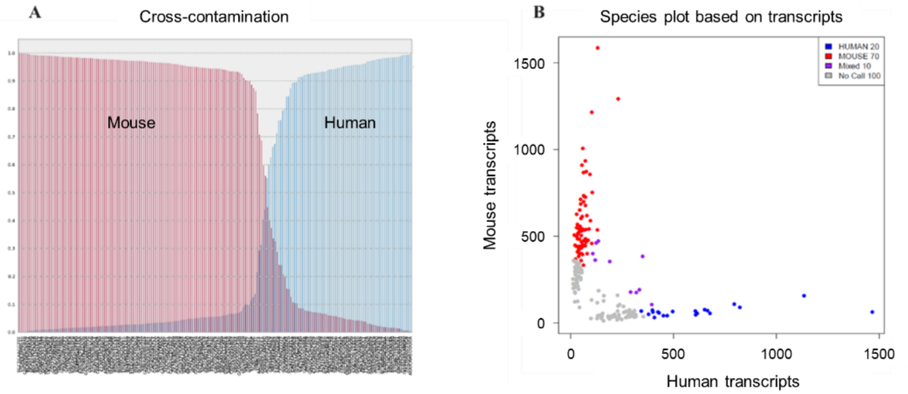


Figure 3.19. Cross-contamination level obtained with modified Drop-seq protocol. **(A)** – the fractions of reads aligned to either human or mouse genomes per cell barcode. Here, the fraction of reads aligned to human genome is shown in blue, and the fraction of reads aligned to mouse genome – in red. **(B)** – the numbers of human and mouse transcripts detected with 100 most abundant cell barcodes.

The proposed approach is in principle compatible with a variety of experimental designs: either with protocols based on reverse transcription in drops with hydrogel beads, such as inDrop (Klein *et al.*, 2015), or those performing reverse transcription in bulk after mRNA capture on hard beads, such as Drop-seq (Macosko *et al.*, 2015). The modified protocol is also applicable to emulsion-free approaches, such as SPLiT-seq (Rosenberg *et al.*, 2018). The only prerequisite is the release of cellular RNA contents within optimal reverse transcription reaction mixture containing oligonucleotide-modified dideoxynucleotides.

4. DISCUSSION

As NGS technologies become increasingly affordable, technology developers around the globe are racing to improve sample preparation methods to maximize the efficiency and informativeness of sequencing. Most current library preparation protocols introduce significant biases, including uneven coverage, artificial chimera, clonal bias caused by amplification, allelic dropouts, length bias in RNA-seq, and more. Several chemoenzymatic library preparation methods have emerged to tackle some of the abovementioned challenges, however such approaches are still in their infancy. Here, we explored the potential of chain terminators bearing bulky oligonucleotide modifications as carriers of sequencing adapters.

4.1 Enzymatic processing of OTDDNs

The idea to employ base-modified chain terminators for nucleic acid analysis is not new, with Sanger sequencing being the closest analog to the approaches developed in this study. Discovery or engineering of polymerases exhibiting low discrimination between dNTPs and ddNTPs played a key role in the improvement of chain termination sequencing performance. T7 DNA polymerase is known to incorporate ddNTPs much more efficiently than wild type DNA polymerases from *E. coli* and *T. aquaticus*, with molecular basis for such activity residing in a single amino acid Y526. Mutations at equivalent positions – F762Y and F667Y in *E. coli* and *Taq* polymerases, respectively – enabled relaxed substrate discrimination and laid the foundation for the engineering of superior enzymes for DNA sequencing (Tabor & Richardson, 1995). This sparked great interest to study sugar recognition mechanisms in other enzymes as well, including *T. litoralis* DNA polymerase, also known as Vent (Gardner & Jack, 1999), DeepVent, 9ºN (Gardner & Jack, 2002), *Pfu* (Evans *et al.*, 2000), *Taq* polymerase Stoffel fragment (Schultz *et al.*, 2015; Chen *et al.*, 2016), and variants of RNA polymerases, such as T7 RNAP (Sousa & Padilla, 1995; Sasaki *et al.*, 1998; Kapustina *et al.*, 2021b) and multi-subunit RNAPs (Mäkinen *et al.*, 2021).

With the emergence of automated terminator sequencing, polymerases encountered a yet more complicated task to incorporate ddNTPs bearing bulky dye modifications. T7 DNA polymerase was found to accept dye-modified dideoxynucleotides consisting of propargylamino linkages to fluorescein dyes (Lee *et al.*, 1992). Studies of *Taq* polymerase variants revealed that the nature of dye modification is important as certain modifications might interfere with a conformational change step which the polymerase undergoes following nucleotide binding (Brandis, 1999). This highlights the importance of careful design of modification as well as conjugation linker. Eventually, sets of highly modified terminators, such as those containing energy-transfer dye labels, were developed and proved to be good substrates for sequencing polymerases like Thermo Sequenase (Kumar & Fuller, 2007). In this study, the molecular design of OTDDNs was found to be compatible with all tested DNA polymerases used in Sanger sequencing – Thermo Sequenase, CycleSeq, Sequenase V2.0 and T7 DNA polymerase – although ON modification is much larger than fluorescent dyes. ON had no apparent adverse effect on polymerases′ discrimination between dNTPs and ddNTPs, e.g. Thermo Sequenase was found to prefer OTDDNs over dNTPs ~2-fold which is consistent with ~2-fold preference for unmodified ddNTPs over dNTPs observed for F667Y *Taq* mutant by Tabor & Richardson, 1995. This allows to expect that any DNA polymerase without proofreading activity able to incorporate ddNTPs would also accept OTDDNs.

Unexpected results were obtained with T7 RNAP mutant V783M able to synthesize transcripts consisting of dTTP, dCTP, dATP and 2-F′-dGTP (Kapustina *et al.*, 2021b). The mutant enzyme was hyper efficient at OTDDN incorporation, exhibiting >1000-fold preference for OTDDNs over dNTPs. The molecular basis of this effect as well as exact roles of sugar and base modification within OTDDN are not clear. It is known that base-modified NTPs are substrates for wild type T7 RNAP (Milisavljevič *et al.*, 2018) but sugar recognition mechanisms of this enzyme should be explored more thoroughly to understand how substrate selection can be modulated.

To sequence RNA molecules, chain termination method was adapted to work with reverse transcriptases (Bauer, 1990). Reverse transcriptases do not possess 3′-5′ exonuclease activity and exhibit relatively low DNA synthesis fidelity which is believed to be linked with retroviral variation (Menéndez-Arias, 2009). The ability to incorporate ddNTPs and general sugar recognition mechanisms are often studied in the context of the development of anti-retroviral therapies (Harris *et al.*, 1998; Selmi *et al.*, 2001). In this study, M-MLV-based RTs as well as HIV RT were able to efficiently utilize OTDDNs as substrates. HIV RT exhibited virtually no discrimination between OTDDNs and dNTPs, which is consistent with more error-prone DNA-dependent cDNA synthesis by HIV RT as compared to M-MLV and AMV RTs (Sebastián-Martín *et al.*, 2018). Structural comparison of HIV RT to M-MLV RT suggested M-MLV residues K103, R110, D153, A154, F155 and Q190 to interact with incoming dNTPs (Halvas *et al.*, 2000). Mutations at K103 and Q190 positions of SuperScript IV reduced the efficiency of OTDDN incorporation, which may reflect more stringent substrate discrimination of these enzyme variants. Interestingly, AMV RT was not able to use OTDDNs as substrates although it is known to incorporate ddNTPs and was used for direct RNA sequencing (Hahn *et al.*, 1989). Probably AMV RT is less tolerant to bulky base modifications or the structure of linker arm within OTDDNs. MarathonRT enzyme, which originates from eubacterial group II intron, was also not able to catalyze OTDDN incorporation, although this activity might have been overlooked because of overall low activity of MarathonRT in PEX assays used in this study. MarathonRT is known for its remarkably high processivity, however it exhibits substantially lower primer utilization efficiency as compared to SuperScript IV (Zhao *et al*., 2018).

Template-independent enzymes investigated in this study – TdT and PUP – both were capable to add a single OTDDN at the 3′ terminus of DNA or RNA, respectively. TdT is well known to tolerate large nucleotide modifications and utilize ddNTPs as substrates as may be exemplified by efficient TdT-mediated addition of streptavidin-modified ddUTP to an ON (Sørensen *et al*., 2013). PUP was less extensively studied in this respect. Examination of Cid 1 PUP crystal structure in the presence of UTP revealed that C5-position of uridine does not contact with neighboring amino acid residues and thus the enzyme might tolerate modifications at this position. This notion was confirmed by efficient RNA labeling with AMUTP, APUTP and ATUTP bearing azide group attached to a nucleobase via increasingly longer linkers with methyl, propyl and tetraethylene glycol spacers, respectively (George *et al*., 2020). Some plasticity in sugar recognition by PUP was observed in the study where azide groups were introduced into 3′ terminus of RNA with 2′-N_3-2′-dNTPs (Winz *et al*., 2012).

A key feature which enabled the development of OTDDN-based nucleic acid analysis applications is the biocompatibility of OTDDN conjugation linker. Triazole backbones were previously shown to be very good DNA backbone analogs (Shivalingam *et al*., 2017). Although the presence of triazole modification affects DNA replication *in vitro* as DNA polymerase either slows down or stalls at modification site, DNA polymerases *Taq*, *Pfu* and Klenow fragment (exo-) were able to read through the modified backbones even when multiple modifications were present (Kukwikila *et al*., 2017). In this study, Klenow fragment (exo-) as well as other exonuclease-deficient enzymes, such as Thermo Sequenase, SuperScript IV and Phusion (exo-), copied DNA strand through the triazole-containing OTDDN linker with varying efficiencies. In our case, read-through efficiency depends not only on enzyme′s tolerance to triazole modification but also on the ability to bypass the linker arm. To explain and predict read-through efficiency of different polymerases, more thorough structural and kinetic studies are needed.

4.2 OTDDNs for DNA sequencing applications

Semi-targeted sequencing approach developed in this work offers an elegant way to study unknown sequence regions nearby the defined loci. Until now, two main technologies employed for this purpose have been inverse PCR and anchored PCR combined with NGS. With the emergence of long-read sequencing technologies (Clarke *et al*., 2009; Eid *et al*., 2009), the studies of structural variation have become more straightforward, although enrichment techniques are still used to increase the coverage of regions of interest (Stangl *et al*., 2020).

Structural variation in which genomic rearrangements act to amplify, delete or reorder chromosomal material at scales ranging from single genes to entire chromosomes is an especially important class of somatic mutations leading to cancer development (Li *et al*., 2020). Long-distance inverse PCR is widely used in cancer research to detect *de novo* DNA rearrangements. Thorsen *et al.* demonstrated the feasibility of this technique to detect *TAF15-ZNF384* and *BCR-ABL1* fusion genes in leukemia samples (Thorsen *et al.*, 2011). Pradhan *et al*. used this assay to screen leiomyoma samples for potential *de novo* breakpoints and identified a novel rearrangement upstream of *HMGA2* (Pradhan *et al*., 2016). Abelleyro and colleagues used inverse PCR for long-distance direct haplotype phasing to improve hemophilia genetic counseling (Abelleyro *et al*., 2019). Inverse PCR procedure includes restriction digestion of genomic DNA and circularization of resulting fragments. Circular molecules are then used as templates for PCR with primers targeting known sequence region and oriented outwards.

Anchored PCR is an alternative approach which employs universal adapter ligation to randomly fragmented dsDNA molecules, and subsequent PCR enrichment using one primer complementary to the adapter and the second primer targeting known fusion partner gene. The technique was proved to be a robust diagnostic assay as well as a discovery tool: new *ARHGEF2-NTRK1* and *CHTOP-NTRK1* gene fusions were identified in glioblastoma, *MSN-ROS1*, *TRIM4-BRAF*, *VAMP2-NRG1*, *TPM3-NTRK1* and *RUFY2-RET* in lung cancer, *FGFR2-CREB5* in cholangiocarcinoma and *PPL-NTRK1* in thyroid carcinoma (Zheng *et al*., 2014). Anchored PCR technology was eventually commercialized by ArcherDX, Inc.

In terms of protocol complexity, semi-targeted sequencing with OTDDNs outperforms both abovementioned methods as no laborious manipulations with sample DNA are needed prior to primer extension. The end result of semi-targeted library preparation is similar to that of anchored PCR. In this study, we looked beyond the traditional applications of probing structural variation and applied semi-targeted sequencing for the characterization of microbial communities. We hypothesized and experimentally proved that direct linking of 16S rRNA gene sequences with regions upstream of the gene enables unambiguous identification of 16S rRNA gene copy numbers as well as more accurate bacteria classification at species level as genomic sequences are less conserved than 16S rRNA gene sequence. In fact, similar reasoning led to the development of RiboFR-seq method by Zhang and colleagues (Zhang *et al*., 2016). RiboFR-seq utilizes inverse PCR to capture regions flanking 16S rRNA gene (Fig. 4.1). Even though the technique was shown to generate meaningful results, the amount of on-target “bridge reads” in sequencing data was <8% which means that the specificity of RiboFR-seq needs serious improvement. In contrast, we obtained >95% of on-target reads by st16S-seq developed in this work.

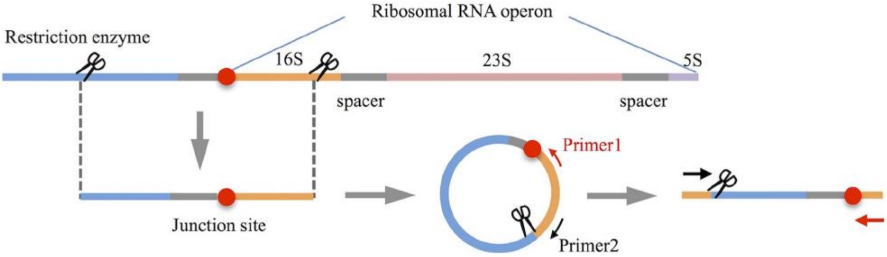


Figure 4.1. Capture of both ribosomal RNA variable regions and their flanking sequences by RiboFR-seq (by Zhang *et al*., 2016).

We thoroughly benchmarked st16S-seq against commercially available PCR-based microbiome characterization techniques with both mock community standards and soil-derived DNA. In our study, amplicons containing 16S rRNA V1-V2 or V1-V3 regions exhibited better performance in terms of species-level classification accuracy and captured alpha diversity in soil samples than those consisting of V4 sequence. In other study involving soil communities (Soriano-Lerma *et al*., 2020) it was observed that V4-V5 domain data clustered separately from all other analyzed 16S rRNA regions in soil samples, indicating that V4-V5 domain was skewed regarding the detection of certain phyla. Our results showed that soil data derived from V4-containing amplicons likewise tended to group together, although we also observed that V1-V3 amplicon formed a distinct cluster. Only st16S-seq and

V1-V2 amplicon datasets clustered along with WGS. Sequencing long-range PCR products, spanning the full-length 16S rRNA gene (Johnson *et al*., 2019; Callahan *et al*., 2019) or 16S-23S rRNA region (Sabat *et al*., 2017), by either short-read or long-read technologies was reported to improve the diagnostic yield in clinical samples. While the rationale to include long-range information lies in capturing greater sequence differences, these techniques are still vulnerable to biases typical for amplicons. Given that st16S-seq requires only one target-specific primer, it is reasonable to believe that robust performance of st16S-seq would depend to a lesser extent on the application and source of bacterial community as is the case with PCR primers (Klindworth *et al*., 2013).

Here, we extensively tested semi-targeted sequencing approach in high-throughput microbiome characterization, however preliminary experiments with more complex templates, such as the capture of *ALK* in human gDNA, provide a promising prospect to develop methods for probing structural variation in human genome. Semi-targeted sequencing design is modular: specific primers can be designed to target any gene of interest and oligonucleotide modification conjugated to dideoxynucleotides can also be of any desired sequence, meaning that the same principle might be adapted for a plethora of applications where highly variable regions are adjacent to defined loci. Moreover, sequencing can be conducted on any platform.

4.3 OTDDNs for RNA sequencing applications

Within the gene expression analysis field, OTDDNs provided means to simplify sample preparation workflows as well as interesting opportunity to sequence cDNA fragments without prior amplification.

PCR-amplified NGS libraries might have reduced complexity compared to the original sample, because different fragments tend to amplify with unequal efficiencies. This in turn causes drop-out of certain transcript species and excessive amplification of others. These pitfalls can be overcome by using UMIs to account for PCR duplicates during data analysis or avoiding library amplification altogether, however conventional library preparation methods are not able to attach sequencing adapters to first-strand cDNA directly. To sequence RNA directly and avoid PCR biases, scientists from the Wellcome Sanger Institute tailored sequencing chemistry to perform reverse transcription on the flow cell, the method was termed flow-cell surface reverse transcription sequencing, or FRT-seq (Mamanova *et al*., 2010). The authors were able to sequence human poly(A)-containing transcripts with 2×37 nt reads. Helicos BioSciences Corporation took over the development of direct

RNA sequencing platform, which feasibility was demonstrated by sequencing *Saccharomyces cerevisiae* RNA with 20-60 nt reads (Ozsolak *et al*., 2009; Lipson *et al*., 2009; Ozsolak & Milos, 2011). Unfortunately, Helicos sequencing system has not achieved commercial success. Here, we used OTDDN coupled to a full-length adapter to label cDNA fragments for loading on a conventional Illumina sequencing machine. This allowed to generate strand-specific transcriptomic data with minimal sample manipulation but retaining good sequencing quality, i.e. efficiency of cluster generation on the flow cell as well as quality of base calling and read lengths standard for Illumina chemistry.

RNA sequencing application which was extensively examined in this study was gene expression analysis by sequencing of mRNA 3′ termini, or MTAS-seq. The closest analog to the method developed herein is Poly(A)-ClickSeq technique, which uses "click" ligation to attach alkyne-functionalized adapter to 3′-azide modified cDNA fragments (Routh *et al*., 2017). Although Poly(A)-ClickSeq procedure is relatively easy to execute, separate chemical ligation step does not allow the development of single-tube protocol and requires an intermediate purification step which inevitably leads to the loss of material. The authors demonstrated that 125 ng of total RNA are minimally required to generate a library, while single-tube MTAS-seq was able to process sub nanogram quantities of starting material. Finally, Poly(A)-ClickSeq generated ~50% of usable reads while >99% of MTAS-seq reads were aligned to the reference genome and processed further. This illustrates the superior technical characteristics of MTAS-seq, retaining all general benefits of 3′ mRNA sequencing approach. Moreover, MTAS-seq was fully functional when whole cell lysates were used instead of purified RNA.

A notable limitation of MTAS-seq emerged upon single-cell RNA sequencing. Although we succeeded to retrieve libraries, transcript capture efficiency was substantially behind that of conventional techniques – more than 20-fold lower than generated by Drop-seq. Systematic benchmarking of available scRNA-seq technologies revealed that Quartz-seq2 (Sasagawa *et al*., 2018) was exceptionally efficient at transcript detection, capturing nearly 6000 genes in HEK293T cells with 20000 reads (Mereu *et al*., 2020). To achieve comparable sensitivity, MTAS-seq protocol should be tailored for ultra-low RNA input amounts. Moreover, the efficiency of DNA synthesis through OTDDN linker should be improved.

For global gene expression profiling, we used oligo(dT) primer at RT step, however modular primer extension and OTDDN labeling design in principle allows to use gene-specific primers as well. This would open up the opportunity to capture specific splice isoforms or fusion transcripts with implications similar to those observed for multiplexed primer extension sequencing (MPE-seq) method (Xu *et al.*, 2019). MPE-seq procedure starts with gene-specific RT primer pool extension in the presence of aminoallyl-dUTP. Resulting cDNAs are then functionalized with biotin and affinity-purified using streptavidin-coated beads. Afterwards, randomly primed first strand extension step generates library of adapter-tagged fragments. MPE-seq was demonstrated to work in a highly multiplexed manner, with nearly 4000 RT primers targeting splice junctions of *Schizosaccharomyces pombe*. Given that OTDDN-based protocols are substantially simpler, we believe that this technology would be successful in targeted RNA-seq applications as well.

Taken together, this work provides a versatile toolbox for high-throughput nucleic acid analysis with a wide spectrum of applications ranging from whole genome and whole transcriptome sequencing to targeted approaches and even exotic possibilities, such as PCR-free RNA sequencing or isothermal synthesis of labeled chimeric nucleic acids. In addition, this work describes substrate and template properties of a novel class of modified chain terminators bearing oligonucleotide modifications and encourages wider adoption of modified nucleotides as tools in genomics research.

CONCLUSIONS

1. C5-substituted 2′,3′-dideoxypyrimidine and C7-substituted 7-deaza-2′,3′-dideoxypurine nucleotide analogs bearing 22-67 nt oligonucleotide modifications attached via triazole-based linker are substrates for Thermo Sequenase, CycleSeq, Sequenase V2.0, T7 DNA polymerase, Phusion (exo-), terminal deoxynucleotidyl transferase, Maxima RT, SuperScript IV RT, SuperScript II RT, RevertAid RT, HIV RT and poly(U) polymerase enzymes.

2. T7 RNA polymerase mutant V783M, engineered towards the reduced substrate discrimination, is able to synthesize transcripts using dTTP, dCTP, dATP, 2′-F-dGTP substrate mixture and incorporate OTDDN to obtain randomly terminated products with labeled 3′ termini.

3. Klenow fragment (exo-), Thermo Sequenase, Phusion (exo-) and SuperScript IV RT are able to synthesize complementary DNA strand through an unnatural linker within oligonucleotide-tethered dideoxynucleotides.

4. Thermo Sequenase labels nascent DNA strand with oligonucleotide-tethered dideoxynucleotides in primer extension reactions in the presence of either random or target-specific primers and ssDNA or dsDNA templates of variable complexity. Resulting DNA fragments constitute a sequenceable library with a characteristic structure: 5′ terminus corresponds to primer sequence while 3′ terminus is dispersed indicating stochastic incorporation of modified dideoxynucleotides.

5. Novel semi-targeted sequencing approach substantially improves microbiome characterization by bridging 16S rRNA gene V1-V2 regions with genomic loci upstream of the gene. New method enables 16S rRNA gene copy number estimation and improves bacteria classification accuracy at species level.

6. SuperScript IV RT labels cDNA with oligonucleotide-tethered dideoxynucleotides in reverse transcription reaction in the presence of either random or oligo(dT) primers and RNA templates. Resulting cDNA fragments constitute sequenceable libraries covering whole transcripts or 3′ termini.

7. Novel 3′ mRNA sequencing method based on oligonucleotide-tethered dideoxynucleotides is feasible for both bulk and single-cell RNA sequencing, and compatible with library preparation from cell lysates without RNA extraction. The technique outperforms conventional methods in terms of specificity to mRNAs, however, underperforms in terms of transcript capture efficiency when RNA inputs are low (<10 ng total RNA).

FUTURE OUTLOOK

Here, the potential of oligonucleotide-modified dideoxynucleotides as tools for high-throughput nucleic acid analysis applications was demonstrated, however there is no doubt that further efforts should be made to optimize proposed nucleic acid labeling methods and create new ones.

A comprehensive study on the mechanisms of OTDDN incorporation and read-through by polymerases would guide enzyme engineering to further improve nucleic acid labeling efficiency. Gene expression analysis applications especially strive for more sensitive transcript capture efficiency, thus pushing the limits of reverse transcription sensitivity as well as the efficiency of amplification of OTDDN-tagged fragments would encourage wider adoption of techniques described herein. Notably, workflow optimization is not a prerogative of enzyme improvement. Further work on the improvement of molecular design of OTDDN linker would also greatly contribute towards the optimization of OTDDN-based methods.

Semi-targeted sequencing might eventually replace traditional targeted amplicon sequencing methods because of substantially higher discovery potential, yet the adoption of this method for highly complex templates requires deeper understanding of oligonucleotide hybridization kinetics. While the overall specificity of PCR is determined by the selectivity of hybridization of both primers and occasional non-specific binding might not reach exponential amplification stage, e.g. if one primer unintendedly hybridizes at a large distance from the other primer, labeling of single primer extension products captures all hybridization events, including imperfect unintended binding which is stable enough to allow the extension of the 3′ terminus. Linear cycles of single primer extension do not contain a specificity compensation mechanism characteristic to exponential PCR, thus new rules for primer design must be created.

Finally, exciting opportunities offered by the attachment of synthetic oligonucleotide to a nucleobase can be further explored not only with dideoxynucleotides, but also with other types of nucleotide analogs. For example, oligonucleotide-tethered cap analogs would enable 5′ end labeling of *in vitro* transcription products, given that oligonucleotide modification is attached via its 3′ terminus and copying enzyme is able to read through the tether linkage in downstream reaction.

Studies in the abovementioned directions will allow to further expand the applicability of base-modified nucleotides and unlock the full potential of chemoenzymatic assays for high-throughput nucleic acid analysis.

SUMMARY/SANTRAUKA

Natūralios nukleotidų modifikacijos, tokios kaip metil, hidroksimetil, formil, karboksi grupės, atlieka svarbų vaidmenį epigenetinėje reguliacijoje (Bilyard *et al.*, 2020). Sintetiniai modifikuoti nukleotidai reikšmingai praplėtė nukleorūgščių pritaikomumo spektrą tokiose srityse kaip terapija, bioanalizė, cheminė biologija, katalizė, biojutikliai, ir kitose (Dhuri *et al.*, 2020; Xu *et al.*, 2017; Lapa *et al.*, 2016; Hollenstein, 2015; Hollenstein *et al.*, 2008). Nukleotidų bazių modifikacijos dažniausiai įvedamos į pirimidinų C5 padėtį arba 7-deazapurinų C7 padėtį (Jäger *et al.*, 2005). Sintetinės modifikacijos gali varijuoti nuo mažų funkcinių grupių iki didelių konjuguotų junginių, tokių kaip baltymai. Gamtinės DNR ir RNR polimerazės pasižymi stebėtinu substrato atpažinimo plastiškumu ir neretai geba naudoti sintetinius nukleotidų analogus kaip substratus. Tai atveria galimybes fermentinei funkcionalizuotų nukleorūgščių sintezei.

Praktinis modifikuotų nukleorūgščių pritaikomumas daugeliu atvejų remiasi jų suderinamumu su fermentinėmis reakcijomis. Nors žinoma, kad fosfodiesteriniai, amidiniai ir triazoliniai DNR karkasai funkcionalūs *in vitro* ir net *in vivo* (Ciafrè *et al.*, 1995; Kuwahara *et al.*, 2009; El-Sagheer *et al.*, 2011; Birts *et al.*, 2014), modifikuotų nukleorūgščių replikacija tiriama sąlyginai mažiau nei nukleotidų analogų įterpimas. Tyrimai rodo, kad fosfatinės grupės pačios savaime nėra būtinos sintetinių matricų kopijavimui, kas leido pritaikyti CuAAC „click“ reakciją fermentinių sistemų toleruojamų modifikuotų nukleorūgščių gamybai. Aukšto tikslumo nenatūralių nukleorūgščių karkasų replikacija įgalintų tirti funkcionalizuotų nukleorūgščių sekas. Pastebėta, jog DNR polimerazės, pasižyminčios klaidų taisymo aktyvumu, sulėtėja ties modifikacijos vieta matricoje, kas veda prie delecijų įvedimo, tuo tarpu fermentai be egzonukleazinio aktyvumo turi mažesnę tikimybę klaidingai prašokti sintetinės modifikacijos sritį (Shivalingam *et al.*, 2017). Modifikuotos nukleorūgštys kartu su suderinamais fermentais nukleotidų analogų įterpimui ir modifikuotų matricų replikacijai sudarytų patrauklų įrankių komplektą naujų molekulinės biologijos metodų vystymui.

Siekdami gerinti mėginių paruošimo naujos kartos sekoskaitai (NKS) procesus bei kurti naujus taikymus mokslininkai žvalgosi į chemofermentinius metodus. „Click“ reakcija, dar vadinama cheminiu ligavimu, leidžia specifiškai prijungti alkino grupe modifikuotą sekoskaitos adapterį prie azido grupe funkcionalizuotų terminatorių, fermentiniu būdu įterptų į sekoskaitai ruošiamas molekules (Miura *et al.*, 2018; Routh *et al.*, 2015). Šis principas turi pranašumų: tinka tiek DNR, tiek kDNR fragmentų bibliotekų paruošimui, pasižymi žemesniu nepageidaujamos rekombinacijos dažniu, taip pat suteikia galimybę ruošti NKS bibliotekas iš viengrandinės DNR. Vis dėlto, vario jonų sąlygota DNR degradacija, žemas konversijos efektyvumas ir daugiastadijiniai protokolai mažina chemofermentinių metodų patrauklumą. Paprasti ir efektyvūs nukleorūgščių žymėjimo sintetiniais oligonukleotidais metodai, nepasižymintys dabartinių chemofermentinių technologijų trūkumais, tačiau išsaugoję jų privalumus, galėtų tapti patraukliu įrankiu nukleorūgščių analizei.

Tikslas ir uždaviniai

Šio darbo **tikslas** buvo ištirti oligonukleotidais modifikuotų 2′,3′-dideoksiribonukleozidų 5′-trifosfatų (ddN^{ON}TP) savybes ir pritaikomumą plataus masto nukleorūgščių analizei, ypatingą dėmesį skiriant DNR ir kDNR molekulių žymėjimui naujos kartos sekoskaita paremtiems taikymams. Tikslui pasiekti buvo iškelti šie **uždaviniai**:

1. Ištirti ddN^{ON}TP, kaip substrato, savybes ir nustatyti fermentus, tinkamus DNR ir RNR žymėjimui.
2. Ištirti fermentų inžinerijos galimybes, siekiant sukurti naujus metodus nukleorūgščių, žymėtų ddN^{ON}TP, sintezei.
3. Ištirti nukleorūgščių, žymėtų ddN^{ON}TP, matricines savybes ir nustatyti DNR polimerazes, gebančias vykdyti sintezę per nenatūralią triazolo žiedą turinčią jungtį.
4. Įvertinti ddN^{ON}TP pritaikomumą DNR ir RNR sekoskaitos taikymams, tiriant tiek visą genomą arba transkriptomą, tiek tikslinius regionus.
5. Pritaikyti DNR žymėjimą ddN^{ON}TP mikrobiomo charakterizavimui.
6. Ištirti kDNR žymėjimo ddN^{ON}TP potencialą genų raiškos analizėje.

Šiame darbe polimerizacijos metu įterpto nukleotido sudėtyje esanti oligonukleotidinė modifikacija buvo panaudota kaip universali pradmens hibridizacijos vieta komplementarios grandinės sintezei, taip sukuriant galimybę specifiškai padauginti šiuo modifikuotu nukleotidu pažymėtus DNR arba kDNR fragmentus. Svarbu ir tai, kad dideoksinukleotidų naudojimas įgalino atsitiktinę sintetinamos DNR grandinės terminaciją suformuojant tinkamo sekoskaitai ilgio fragmentus.

Darbo metu parodyta, kad efektyvų fermentinį ddN^{ON}TP įterpimą vykdo Thermo Sequenase, CycleSeq, Sequenase V2.0 ir TdT polimerazės, o kopijuojant iRNR sintetinama kDNR gali būti efektyviai pažymėta naudojant Maxima, SuperScript IV, SuperScript II, RevertAid ir ŽIV atvirkštines transkriptazes. RNR žymėjimui tinka poliU polimerazė, tuo tarpu šiame darbe *in vitro* evoliucijos pagalba atrinktas T7 RNR polimerazės mutantas V783M, gebantis sintetinti iš dTTP, dCTP, dATP ir 2′-F-dGTP sudarytus vgDNR transkriptus, taip pat geba įjungti ddN^{ON}TP atsitiktinėse transkripto vietose. Tai leidžia pažymėti transkripcijos produktų 3′ galus universalia seka, o tuo pačiu leidžia linijiškai padauginti NKS tinkamas DNR sekas – tai yra ypač aktualu dirbant su itin mažais DNR kiekiais. Nustatyta, kad triazolo žiedą turinti netipinė jungtis tarp bazės ir prie jos prijungto oligonukleotido (ddN^{ON}TP sudėtyje) komplementarios grandinės sintezės metu toleruojama Phusion exo-, Klenow fragmento exo-, Thermo Sequenase ir SuperScript IV DNR polimerazių, iš kurių Phusion exo- pasižymėjo geriausiu perskaitymo efektyvumu. Sėkmingas įterpimo ir netipinės jungties perskaitymo fermentų identifikavimas įgalino pritaikyti ddN^{ON}TP technologiją fragmentų bibliotekų paruošimui naujos kartos sekoskaitai. Be to, parodyta, kad oligonukleotidinė modifikacija gali turėti randomizuotos sekos regionus arba afininius ligandus, kurie vėliau tarnauja atitinkamai kaip molekuliniai barkodai arba afininiai žymenys patogiam tikslinių molekulių gryninimui.

Atsitiktinis ddN^{ON}TP įterpimo pobūdis ir tuo pat metu vykstantis sintetinamos grandinės žymėjimas universalia seka tapo naujo sekoskaitos bibliotekų konstravimo metodo vystymo pagrindu. Parodyta, kad atsitiktinių pradmenų pratęsimo produktų žymėjimas leidžia paruošti visą genomą arba visą transkriptomą dengiančias NKS bibliotekas, o tuo tarpu specifinių pradmenų pratęsimas ir žymėjimas konstruojant pusiau taikinines (*angl.* semi-targeted) bibliotekas leidžia tirti nežinomos sekos regionus šalia pasirinktų lokusų. Įrodyta, kad ši strategija yra naudinga mikrobiomo charakterizavimui – sukurtas naujas st16S-seq metodas, įgalinęs tirti *a priori* nežinomus genominius lokusus, esančius prieš 16S rRNR geną. Naujas metodas turi rimtų pranašumų prieš tradicinius: tikslus 16S rRNR geno kopijų nustatymas, tikslesnis bakterinių rūšių charakterizavimas ir mažesnė metodo priklausomybė nuo pradmenų dizaino, kadangi st16S-seq metodui reikia tik į vieną pusę orientuotų specifinių pradmenų.

RNR sekoskaitos srityje buvo sukurtas naujas genų raiškos analizės metodas MTAS-seq, kuris įgalino ženkliai supaprastinti mėginio paruošimo procesą ir užtikrino aukštos kokybės NKS duomenis dirbant tiek su išgryninta RNR, tiek ir ruošiant mėginius tiesiogiai eukariotinių ląstelių lizatuose. Be to, 3′UTR regionų praturtinimas leidžia analizuoti alternatyvaus poliadenilinimo profilius eukariotiniuose transkriptomuose. Įdomu tai, kad ddN^{ON}TP su modifikacija, atitinkančia pilno ilgio sekoskaitos adapterį, įgalino sekvenuoti kDNR fragmentus be PGR amplifikacijos, kas anksčiau nebuvo įmanoma su standartinėmis NKS sistemomis.

Šis darbas siūlo ddN^{ON}TP kaip perspektyvų įrankį NKS taikymams, kuris leidžia kurti tiek patogesnius protokolus, tiek ir visiškai naujus nukleorūgščių analizės metodus. Suformuluotos darbo **išvados**:

1. 5-pakeisti 2′,3′-dideoksipirimidinų ir 7-deaza-7-pakeisti-2′,3′-dideoksipurinų nukleotidų analogai, turintys 22-67 nt oligonukleotidines modifikacijas, konjuguotas per triazolo žiedą turinčią jungtį, yra substratai Thermo Sequenase, CycleSeq, Sequenase V2.0, T7 DNR, Phusion (exo-), TdT polimerazėms, Maxima, SuperScript IV, SuperScript II, RevertAid, ŽIV atvirkštinėms transkriptazėms bei poliU RNR polimerazei.
2. *In vitro* evoliucijos būdu identifikuotas T7 RNR polimerazės mutantas V783M, gebantis sintetinti transkriptus iš dTTP, dCTP, dATP ir 2′-F-dGTP bei įterpinėti ddN^{ON}TP, generuojant atsitiktinai terminuotus transkripcijos produktus, pažymėtus oligonukleotidu 3′ gale.
3. Klenow fragmentas (exo-), Thermo Sequenase, Phusion (exo-) ir SuperScript IV geba sintetinti komplementarią DNR grandinę per nenatūralią ddN^{ON}TP jungtį.
4. Thermo Sequenase fermentas pažymi sintetinamas DNR grandines ddN^{ON}TP pradmenų pratęsimo reakcijose esant pradmenims ir vgDNR arba dgDNR matricoms. Gaunami fragmentai sudaro charakteringos struktūros sekoskaitai paruoštą biblioteką: fragmentų 5′ galai atitinka pratęsimui naudotų pradmenų sekas, o 3′ galų pozicijos pasiskirsčiusios atsitiktinai dėl stochastinio modifikuotų dideoksinukleotidų įterpimo pobūdžio.
5. Naujas pusiau taikininės sekoskaitos metodas, sujungęs 16S rRNR geno V1-V2 regionų sekas su genominėmis sekomis, esančiomis prieš 16S rRNR geną, reikšmingai pagerino mikrobiomo charakterizavimą. Naujas metodas įgalina nustatyti 16S rRNR geno kopijų skaičių bei reikšmingai pagerina klasifikacijos tikslumą rūšių lygmenyje.
6. SuperScript IV pažymi sintetinamas kDNR grandines ddN^{ON}TP atvirkštinės transkripcijos reakcijoje esant pradmenims ir RNR matricoms. Gaunami kDNR fragmentai sudaro sekoskaitai tinkamą biblioteką, praturtintą 3′ UTR/poliA regionais arba dengiančią visą transkripto ilgį, priklausomai nuo naudotų pradmenų sekų.
7. Naujas 3′ iRNR sekoskaitos metodas paremtas ddN^{ON}TP žymėjimu tinka tiek tradicinei RNR sekoskaitai, tiek pavienių ląstelių analizei. Naujas metodas lenkia tradicinius bibliotekų ruošimo būdus specifiškumu iRNR, tačiau nusileidžia jautrumu, kuomet mėginyje RNR kiekis labai mažas (<10 ng suminės RNR).

SCIENTIFIC PARTICIPATION

Patent applications

- A. Lubys, I. Čikotienė, **Ž. Kapustina**, A. Berezniakovas, J. Medžiūnė, S. Žeimytė. Tethered Oligos And Uses Thereof. International Application No. PCT/US2020/039009.
- A. Lubys, **Ž. Kapustina**, A. Jasponė. Mutant Polymerases And Methods Of Using The Same. USPTO Application No. 63/023,026.
- A. Lubys, A. Grybauskas, D. Strepetkaitė, **Ž. Kapustina**, A. Markina, P. Mielinis. Isolated Nucleic Acid Binding Domains. International Application No. PCT/EP2020/061656.

Publications

Research articles directly related to the topic of doctoral dissertation, published in journals with an impact factor (IF) in the Clarivate Analytics Web of Science platform:

- **Kapustina Ž**, Medžiūnė J, Alzbutas G, Rokaitis I, Matjošaitis K, Mackevičius G, Žeimytė S, Karpus L, Lubys A. High-resolution microbiome analysis enabled by linking of 16S rRNA gene sequences with adjacent genomic contexts. *Microb Genom.* 2021a;7(9). doi: 10.1099/mgen.0.000624.
- **Kapustina Ž**, Jasponė A, Dubovskaja V, Mackevičius G, Lubys A. Enzymatic Synthesis of Chimeric DNA Oligonucleotides by *in Vitro* Transcription with dTTP, dCTP, dATP, and 2'-Fluoro Modified dGTP. *ACS Synth Biol.* 2021b;10(7):1625-1632.
- Medžiūnė J, **Kapustina Ž**, Žeimytė S, Jakubovska J, Sindikevičienė R, Čikotienė I, Lubys A. Advanced preparation of fragment libraries enabled by oligonucleotide-modified 2′,3′-dideoxynucleotides. Manuscript in review.

Publications which are not directly related to the topic of doctoral dissertation in journals with an impact factor (IF) in the Clarivate Analytics Web of Science platform:

- Gasiulė S, Dreižė N, Kaupinis A, Ražanskas R, Čiupas L, Stankevičius V, **Kapustina Ž**, Laurinavičius A, Valius M, Vilkaitis G. miR-23b Modulates the Epithelial-Mesenchymal Transition of Colorectal Cancer Cells. *J Clin Med*. 2019;8(12).

- Gasiulė S, Stankevičius V, Patamsytė V, Ražanskas R, Žukovas G, **Kapustina Ž**, Žaliaduonytė D, Benetis R, Lesauskaitė V, Vilkaitis G. Tissue-Specific miRNAs Regulate the Development of Thoracic Aortic Aneurysm: The Emerging Role of KLF4 Network. *J Clin Med*. 2019;8(10).

Conference presentations

Poster presentations directly related to the topic of doctoral dissertation:

- **Kapustina Ž**, Šulgaitė J, Žeimytė S, Lubys A. Innovative approach for digital gene expression and alternative polyadenylation profiling. European Human Genetics Virtual Conference, ESHG 2020.
- **Kapustina Ž**, Medžiūnė J, Žeimytė S, Liucvaikytė A, Lubys A. Innovative approach for digital gene expression and alternative polyadenylation profiling. American Society of Human Genetics Virtual Meeting, ASHG 2020.
- **Kapustina Ž**, Alzbutas G, Šulgaitė J, Žeimytė S, Lubys A. High-precision characterization of microbial communities by the analysis of 16S rRNA gene genomic context. Cold Spring Harbor Laboratory Microbiome Virtual Meeting, 2020.
- **Kapustina Ž**, Medžiūnė J, Drazdauskienė U, Sindikevičienė R, Dubovskaja V, Lubys A. Exploratory analysis of genetic rearrangements. 2021 Advances in Genome Biology and Technology Virtual General Meeting, AGBT 2021.
- Jakubovska J, Sindikevičienė R, Medžiūnė J, Kerzhner M, Kuhn P, **Kapustina Ž**, Lubys A. Oligonucleotide-modified terminators for high-throughput single-stranded DNA sequencing. Revolutionizing Next-Generation Sequencing (Virtual 4th edition), VIB Tools and Technologies conference 2021.
- Drazdauskienė U, **Kapustina Ž**, Medžiūnė J, Gasiūnienė M, Sindikevičienė R, Dubovskaja V, Sabaliauskaitė R, Lubys A. Semi-targeted sequencing of fusion transcripts in prostate cancer. European Human Genetics Conference, ESHG 2021.

Poster presentations prepared during doctoral studies but not related to the topic of doctoral dissertation:

- **Kapustina Ž**, Lubienė J, Alzbutas G, Lubys A. High-throughput genome-wide analysis of *Escherichia coli* essential genes. 14th International Conference of Lithuanian Biochemical Society, 2016, Druskininkai, Lithuania.
- **Kapustina Ž**, Alzbutas G, Lubys A. Phenotypic screening of *Escherichia coli* mutants exhibiting altered transcriptional profiles. 7th Congress of European Microbiologists, FEMS, 2017, Valencia, Spain.
- **Kapustina Ž**, Alzbutas G, Lubienė J, Lubys A. Next-generation *Escherichia coli* strains. 15th International Conference of Lithuanian Biochemical Society, 2018, Dubingiai, Lithuania.

CURRICULUM VITAE

Name, Surname	Žana Kapustina
General information	
Date of birth	20th December, 1991
Main workplace	Thermo Fisher Scientific Baltics, UAB V.A.Graičiūno str. 8, Vilnius 02241 Research and Development Advanced technology group
Personal contacts	Phone: +37068371332 Email: zana.kapustina@thermofisher.com zana@olimpiados.lt
Education	
2016 – 2020	Doctoral studies (Biology, N 010); Institute of Biosciences, Life Sciences Center, Vilnius University
2014 – 2016	MSc Genetics (*Magna Cum Laude*), Life Sciences Center, Vilnius University
2010 – 2014	BSc Genetics (*Cum Laude*), former Faculty of Natural Sciences, Vilnius University
2008 – 2010	Klaipėda Ąžuolynas Gymnasium
1998 – 2008	Klaipėda Maxim Gorky School
Training courses	
2020	„Manager´s start“ by ISM Executive School, Vilnius, Lithuania
2018	„Single-cell analysis“ by TATAA Biocenter at Single Cell Europe 2018, Prague, Czech Republic
2012	„Real-time quantitative PCR“, „Sample preparation and quality control of nucleic acids“, „Quality control of qPCR in molecular diagnostics“, „Genotyping with qPCR“ by TATAA Biocenter, Gothenburg, Sweden
Work experience	
Since 2020	R&D Manager at Thermo Fisher Scientific Baltics, UAB
2019 – 2020	Scientist III at Thermo Fisher Scientific Baltics, UAB
2018 – 2019	Scientist II at Thermo Fisher Scientific Baltics, UAB
2015 – 2018	Junior Scientist at Thermo Fisher Scientific Baltics, UAB
2012 – 2014	Laboratory assistant at former Faculty of Natural Sciences, Vilnius University
Teaching experience	
2018 – 2021	Molecular genetics methods (VU Genetics MSc program) Transcriptomics (VU Systems biology MSc program)

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor prof. dr. Arvydas Lubys for the provided opportunity to perform doctoral research and start my career at Thermo Fisher Scientific Baltics R&D. Thank you for all the challenges, fruitful discussions, valuable advices, critical comments and great mentorship throughout this journey. I admit I was not easy to work with sometimes, thank you for your patience and support.

I am grateful to my colleague Justina Medžiūnė who developed and executed the synthesis of oligonucleotide-tethered dideoxynucleotides. Thank you for your flexibility and openness to new challenges.

I sincerely appreciate the great work on T7 RNA polymerase development executed by Aistė Jasponė. Thank you for hard work and persistence, and for all your moral support and understanding.

Warmest thanks to dr. Gediminas Alzbutas, dr. Varvara Dubovskaja, Karolis Matjošaitis and Gytis Mackevičius for their excellent support with NGS data analysis and the development of specialized analysis pipelines.

I thank Judita Lubienė for guidance at the very beginning of my time at Thermo Fisher Scientific. Moreover, I am grateful to all colleagues from Sequencing and Protein *in vitro* evolution labs led by dr. Vaida Šeputienė and dr. Remigijus Skirgaila, respectively, for constant support. Special thanks to all former and present members of the Advanced technology group – I am truly lucky to work with enthusiastic and ambitious scientists; you have inspired me to strive for more and finally finish my doctoral studies.

Thank you to all my students – Karolina, Aistė, Eglė, Jenni, Mindaugas, Laura and Ugnė – who delighted me with their achievements and enabled me to improve not only as a scientist but also as a mentor and a person.

Furthermore, I am grateful to Elena Armalytė, Indrė Jackūnė, dr. Eglė Radzevičienė and other members of the Intellectual property group for their excellent work on our patent applications as well as their help, support and valuable comments on all of my conference papers and manuscripts.

Finally, warmest gratitude to my family and friends who believe in me no matter the circumstances.

REFERENCES

1. Abelleyro MM, Marchione VD, Palmitelli M, Radic CP, Neme D, Larripa IB, Medina-Acosta E, De Brasi CD, Rossetti LC. Inverse PCR to perform long-distance haplotyping: main applications to improve preimplantation genetic diagnosis in hemophilia. *Eur J Hum Genet*. 2019;27(4):603-611.
2. Aigrain L, Gu Y, Quail MA. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing. *BMC Genomics*. 2016;17:458.
3. Anderson JP, Angerer B, Loeb LA. Incorporation of reporter-labeled nucleotides by DNA polymerases. *Biotechniques*. 2005;38(2):257-64.
4. Anhäuser L, Rentmeister A. Enzyme-mediated tagging of RNA. *Curr Opin Biotechnol*. 2017;48:69-76.
5. Arias-Gonzalez JR. A DNA-centered explanation of the DNA polymerase translocation mechanism. *Sci Rep*. 2017;7(1):7566.
6. Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM, Raymond CK. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods.* 2009;6(9):647-9.
7. Asare-Okai PN, Agustin E, Fabris D, Royzen M. Site-specific fluorescence labelling of RNA using bio-orthogonal reaction of trans-cyclooctene and tetrazine. *Chem Commun (Camb).* 2014;50(58):7844-7.
8. Baccaro A, Marx A. Enzymatic synthesis of organic-polymer-grafted DNA. *Chemistry*. 2010;16(1):218-26.
9. Baccaro A, Steck AL, Marx A. Barcoded nucleotides. *Angew Chem Int Ed Engl.* 2012;51(1):254-7.
10. Balintová J, Welter M, Marx A. Antibody-nucleotide conjugate as a substrate for DNA polymerases. *Chem Sci*. 2018;9(35):7122-7125.
11. Bauer GJ. RNA sequencing using fluorescent-labeled dideoxynucleotides and automated fluorescence detection. *Nucleic Acids Res.* 1990;18(4):879-84.
12. Bergen K, Steck AL, Strütt S, Baccaro A, Welte W, Diederichs K, Marx A. Structures of KlenTaq DNA polymerase caught while incorporating C5-modified pyrimidine and C7-modified 7-deazapurine nucleoside triphosphates. *J Am Chem Soc*. 2012;134(29):11840-3.
13. Beyer C, Wagenknecht HA. *In situ* azide formation and "click" reaction of nile red with DNA as an alternative postsynthetic route. *Chem Commun (Camb).* 2010;46(13):2230-1.

14. Bilyard MK, Becker S, Balasubramanian S. Natural, modified DNA bases. *Curr Opin Chem Biol*. 2020;57:1-7.
15. Birts CN, Sanzone AP, El-Sagheer AH, Blaydes JP, Brown T, Tavassoli A. Transcription of click-linked DNA in human cells. *Angew Chem Int Ed Engl*. 2014;53(9):2362-5.
16. Brandis JW. Dye structure affects Taq DNA polymerase terminator selectivity. *Nucleic Acids Res*. 1999;27(8):1912-8.
17. Buermans HP, den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta*. 2014;1842(10):1932-1941.
18. Buhr CA, Wagner RW, Grant D, Froehler BC. Oligodeoxynucleotides containing C-7 propyne analogs of 7-deaza-2'-deoxyguanosine and 7-deaza-2'-deoxyadenosine. *Nucleic Acids Res*. 1996;24(15):2974-80.
19. Cahová H, Pohl R, Bednárová L, Nováková K, Cvacka J, Hocek M. Synthesis of 8-bromo-, 8-methyl- and 8-phenyl-dATP and their polymerase incorporation into DNA. *Org Biomol Chem*. 2008;6(20):3657-60.
20. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res*. 2019;47(18):e103.
21. Carell T, Brandmayr C, Hienzsch A, Müller M, Pearson D, Reiter V, Thoma I, Thumbs P, Wagner M. Structure and function of noncanonical nucleobases. *Angew Chem Int Ed Engl*. 2012;51(29):7110-31.
22. Chen B, Jiang L, Zhong ML, Li JF, Li BS, Peng LJ, Dai YT, Cui BW, Yan TQ, Zhang WN, Weng XQ, Xie YY, Lu J, Ren RB, Chen SN, Hu JD, Wu DP, Chen Z, Tang JY, Huang JY, Mi JQ, Chen SJ. Identification of fusion genes and characterization of transcriptome features in T-cell acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A*. 2018;115(2):373-378.
23. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, Xie XS. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science*. 2017;356(6334):189-194.
24. Chen S, Qiu G, Yang M. SMRT sequencing of full-length transcriptome of seagrasses *Zostera japonica. Sci Rep*. 2019;9(1):14537.
25. Chen T, Hongdilokkul N, Liu Z, Adhikary R, Tsuen SS, Romesberg FE. Evolution of thermophilic DNA polymerases for the recognition and amplification of C2'-modified DNA. *Nat Chem*. 2016;8(6):556-62.
26. Ciafrè SA, Rinaldi M, Gasparini P, Seripa D, Bisceglia L, Zelante L, Farace MG, Fazio VM. Stability and functional effectiveness of

phosphorothioate modified duplex DNA and synthetic 'mini-genes'. *Nucleic Acids Res*. 1995;23(20):4134-42.

27. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*. 2009;4(4):265-70.
28. Curanovic D, Cohen M, Singh I, Slagle CE, Leslie CS, Jaffrey SR. Global profiling of stimulus-induced polyadenylation in cells using a poly(A) trap. *Nat Chem Biol*. 2013;9(11):671-3.
29. D'Argenio V, Esposito MV, Telese A, Precone V, Starnone F, Nunziato M, Cantiello P, Iorio M, Evangelista E, D'Aiuto M, Calabrese A, Frisso G, D'Aiuto G, Salvatore F. The molecular analysis of *BRCA1* and *BRCA2*: Next-generation sequencing supersedes conventional approaches. *Clin Chim Acta*. 2015;446:221-5.
30. D'Argenio V. The High-Throughput Analyses Era: Are We Ready for the Data Struggle? *High Throughput*. 2018;7(1):8.
31. Das D, Georgiadis MM. The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. *Structure*. 2004;12(5):819-29.
32. Dellafiore MA, Montserrat JM, Iribarren AM. Modified Nucleoside Triphosphates for *In-vitro* Selection Techniques. *Front Chem*. 2016;4:18.
33. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM. Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol*. 2011;77(4):1315-24.
34. Dhuri K, Bechtold C, Quijano E, Pham H, Gupta A, Vikram A, Bahal R. Antisense Oligonucleotides: An Emerging Area in Drug Discovery and Development. *J Clin Med*. 2020;9(6):2004.
35. Di Giusto DA, Wlassoff WA, Giesebrecht S, Gooding JJ, King GC. Enzymatic synthesis of redox-labeled RNA and dual-potential detection at DNA-modified electrodes. *Angew Chem Int Ed Engl*. 2004;43(21):2809-12.
36. Duan B, Wu S, Da LT, Yu J. A critical residue selectively recruits nucleotides for T7 RNA polymerase transcription fidelity control. *Biophys J*. 2014;107(9):2130-40.
37. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin

D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133-8.

38. Ellington AD, Szostak JW. Selection *in vitro* of single-stranded DNA molecules that fold into specific ligand-binding structures. *Nature*. 1992;355(6363):850-2.

39. Elrod ND, Jaworski EA, Ji P, Wagner EJ, Routh A. Development of Poly(A)-ClickSeq as a tool enabling simultaneous genome-wide poly(A)-site identification and differential expression analysis. *Methods*. 2019;155:20-29.

40. El-Sagheer AH, Sanzone AP, Gao R, Tavassoli A, Brown T. Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 2011;108(28):11338-43.

41. Espejo RT, Plaza N. Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA. *Front Microbiol.* 2018;9:1232.

42. Evans SJ, Fogg MJ, Mamone A, Davis M, Pearl LH, Connolly BA. Improving dideoxynucleotide-triphosphate utilisation by the hyper-thermophilic DNA polymerase from the archaeon *Pyrococcus furiosus*. *Nucleic Acids Res*. 2000;28(5):1059-66.

43. Freier SM, Altmann KH. The ups and downs of nucleic acid duplex stability: structure-stability studies on chemically-modified DNA:RNA duplexes. *Nucleic Acids Res*. 1997;25(22):4429-43.

44. Gansauge MT, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat Protoc*. 2013;8(4):737-48.

45. Gao L, Fang Z, Zhang K, Zhi D, Cui X. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*. 2011;27(5):662-9.

46. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods.* 2018;15(3):201-206.

47. Gardner AF, Jack WE. Acyclic and dideoxy terminator preferences denote divergent sugar recognition by archaeon and Taq DNA polymerases. *Nucleic Acids Res*. 2002;30(2):605-13.

48. Gardner AF, Jack WE. Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res*. 1999;27(12):2545-53.

49. George JT, Azhar M, Aich M, Sinha D, Ambi UB, Maiti S, Chakraborty D, Srivatsan SG. Terminal Uridylyl Transferase Mediated Site-Directed

Access to Clickable Chromatin Employing CRISPR-dCas9. *J Am Chem Soc*. 2020;142(32):13954-13965.

50. George JT, Srivatsan SG. Vinyluridine as a Versatile Chemoselective Handle for the Post-transcriptional Chemical Functionalization of RNA. *Bioconjug Chem*. 2017;28(5):1529-1536.
51. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, Carter J, Dalby AB, Eaton BE, Fitzwater T, Flather D, Forbes A, Foreman T, Fowler C, Gawande B, Goss M, Gunn M, Gupta S, Halladay D, Heil J, Heilig J, Hicke B, Husar G, Janjic N, Jarvis T, Jennings S, Katilius E, Keeney TR, Kim N, Koch TH, Kraemer S, Kroiss L, Le N, Levine D, Lindsey W, Lollo B, Mayfield W, Mehan M, Mehler R, Nelson SK, Nelson M, Nieuwlandt D, Nikrad M, Ochsner U, Ostroff RM, Otis M, Parker T, Pietrasiewicz S, Resnicow DI, Rohloff J, Sanders G, Sattin S, Schneider D, Singer B, Stanton M, Sterkel A, Stewart A, Stratford S, Vaught JD, Vrkljan M, Walker JJ, Watrobka M, Waugh S, Weiss A, Wilcox SK, Wolfson A, Wolk SK, Zhang C, Zichi D. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One*. 2010;5(12):e15004.
52. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Sano Marma M, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J. Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A*. 2008;105(27):9145-50.
53. Hahn CS, Strauss EG, Strauss JH. Dideoxy sequencing of RNA using reverse transcriptase. *Methods Enzymol*. 1989;180:121-30.
54. Halvas EK, Svarovskaia ES, Pathak VK. Role of murine leukemia virus reverse transcriptase deoxyribonucleoside triphosphate-binding site in retroviral replication and *in vivo* fidelity. *J Virol*. 2000;74(22):10349-58.
55. Harris D, Kaushik N, Pandey PK, Yadav PN, Pandey VN. Functional analysis of amino acid residues constituting the dNTP binding pocket of HIV-1 reverse transcriptase. *J Biol Chem*. 1998;273(50):33624-34.
56. Hashimoto JG, Stevenson BS, Schmidt TM. Rates and consequences of recombination between rRNA operons. *J Bacteriol*. 2003;185(3):966-72.
57. Hatahet Z, Purmal AA, Wallace SS. A novel method for site specific introduction of single model oxidative DNA lesions into oligodeoxyribonucleotides. *Nucleic Acids Res*. 1993;21(7):1563-8.
58. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 2014;56(2):61-4, 66, 68, passim.

59. Hocek M, Fojta M. Cross-coupling reactions of nucleoside triphosphates followed by polymerase incorporation. Construction and applications of base-functionalized nucleic acids. *Org Biomol Chem*. 2008;6(13):2233-41.
60. Hocek M. Enzymatic Synthesis of Base-Functionalized Nucleic Acids for Sensing, Cross-linking, and Modulation of Protein-DNA Binding and Transcription. *Acc Chem Res*. 2019;52(6):1730-1737.
61. Hocek M. Synthesis of base-modified 2'-deoxyribonucleoside triphosphates and their use in enzymatic synthesis of modified DNA for applications in bioanalysis and chemical biology. *J Org Chem*. 2014;79(21):9914-21.
62. Hollenstein M, Hipolito C, Lam C, Dietrich D, Perrin DM. A highly selective DNAzyme sensor for mercuric ions. *Angew Chem Int Ed Engl*. 2008;47(23):4346-50.
63. Hollenstein M. DNA Catalysis: The Chemical Repertoire of DNAzymes. *Molecules*. 2015;20(11):20777-804.
64. Hollenstein M. Nucleoside triphosphates--building blocks for the modification of nucleic acids. *Molecules*. 2012;17(11):13569-91.
65. Hollenstein M. Synthesis of deoxynucleoside triphosphates that include proline, urea, or sulfonamide groups and their polymerase incorporation into DNA. *Chemistry*. 2012 Oct;18(42):13320-30.
66. Horiya S, MacPherson IS, Krauss IJ. Recent strategies targeting HIV glycans in vaccine design. *Nat Chem Biol*. 2014;10(12):990-9.
67. Hoshika S, Leal NA, Kim MJ, Kim MS, Karalkar NB, Kim HJ, Bates AM, Watkins NE Jr, SantaLucia HA, Meyer AJ, DasGupta S, Piccirilli JA, Ellington AD, SantaLucia J Jr, Georgiadis MM, Benner SA. Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science*. 2019;363(6429):884-887.
68. Hottin A, Betz K, Diederichs K, Marx A. Structural Basis for the KlenTaq DNA Polymerase Catalysed Incorporation of Alkene- versus Alkyne-Modified Nucleotides. *Chemistry*. 2017;23(9):2109-2118.
69. Hottin A, Marx A. Structural Insights into the Processing of Nucleobase-Modified Nucleotides by DNA Polymerases. *Acc Chem Res*. 2016;49(3):418-27.
70. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA.* 2017;8(1):10.1002/wrna.1364.
71. Huang CE, Ma GC, Jou HJ, Lin WH, Lee DJ, Lin YS, Ginsberg NA, Chen HF, Chang FM, Chen M. Noninvasive prenatal diagnosis of fetal aneuploidy by circulating fetal nucleated red blood cells and extravillous trophoblasts using silicon-based nanostructured microfluidics. *Mol Cytogenet.* 2017;10:44.

72. Huptas C, Scherer S, Wenning M. Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC Res Notes*. 2016;9:269.
73. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50(8):96.
74. Ilott NE, Ponting CP. Predicting long non-coding RNAs using RNA sequencing. *Methods*. 2013;63(1):50-9.
75. Ingale SA, Pujari SS, Sirivolu VR, Ding P, Xiong H, Mei H, Seela F. 7-Deazapurine and 8-aza-7-deazapurine nucleoside and oligonucleotide pyrene "click" conjugates: synthesis, nucleobase controlled fluorescence quenching, and duplex stability. *J Org Chem*. 2012;77(1):188-99.
76. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc.* 2012;7(5):813-28.
77. Ivancová I, Leone DL, Hocek M. Reactive modifications of DNA nucleobases for labelling, bioconjugations, and cross-linking. *Curr Opin Chem Biol.* 2019;52:136-144.
78. Jäger S, Rasched G, Kornreich-Leshem H, Engeser M, Thum O, Famulok M. A versatile toolbox for variable DNA functionalization at high density. *J Am Chem Soc*. 2005;127(43):15071-82.
79. Jakubovska J, Tauraite D, Birštonas L, Meškys R. *N*^{4}-acyl-2'-deoxycytidine-5'-triphosphates for the enzymatic synthesis of modified DNA. *Nucleic Acids Res*. 2018;46(12):5911-5923.
80. Jensen KB, Atkinson BL, Willis MC, Koch TH, Gold L. Using *in vitro* selection to direct the covalent attachment of human immunodeficiency virus type 1 Rev protein to high-affinity RNA ligands. *Proc Natl Acad Sci U S A*. 1995;92(26):12220-4.
81. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun*. 2019;10(1):5029.
82. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A*. 2006;103(52):19635-40.
83. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* 2015;43(21):e143.
84. Kennedy WP, Momand JR, Yin YW. Mechanism for *de novo* RNA synthesis and initiating nucleotide specificity by T7 RNA polymerase. *J Mol Biol.* 2007;370(2):256-68.

85. Kircher M, Heyn P, Kelso J. Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics*. 2011;12:382.
86. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161(5):1187-1201.
87. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. 2013;41(1):e1.
88. Knapp DC, Serva S, D'Onofrio J, Keller A, Lubys A, Kurg A, Remm M, Engels JW. Fluoride-cleavable, fluorescently labelled reversible terminators: synthesis and use in primer extension. *Chemistry*. 2011;17(10):2903-15.
89. Koh HR, Roy R, Sorokina M, Tang GQ, Nandakumar D, Patel SS, Ha T. Correlating Transcription Initiation and Conformational Changes by a Single-Subunit RNA Polymerase with Near Base-Pair Resolution. *Mol Cell*. 2018;70(4):695-706.e5.
90. Kovacs T, Otvos L. Simple synthesis of 5-vinyl- and 5-ethynyl-2′deoxyuridine-5′-triphosphates. *Tetrahedron Lett*. 1988;29:4525–28.
91. Kropp HM, Dürr SL, Peter C, Diederichs K, Marx A. Snapshots of a modified nucleotide moving through the confines of a DNA polymerase. *Proc Natl Acad Sci U S A*. 2018;115(40):9992-9997.
92. Kukwikila M, Gale N, El-Sagheer AH, Brown T, Tavassoli A. Assembly of a biocompatible triazole-linked gene by one-pot click-DNA ligation. *Nat Chem*. 2017;9(11):1089-1098.
93. Kumar S, Fuller C. Advances in Dye-Nucleotide Conjugate Chemistry for DNA Sequencing. *Perspectives in Bioanalysis*. 2007;2:119-149.
94. Kuwahara M, Nagashima J, Hasegawa M, Tamura T, Kitagata R, Hanawa K, Hososhima S, Kasamatsu T, Ozaki H, Sawai H. Systematic characterization of 2'-deoxynucleoside- 5'-triphosphate analogs as substrates for DNA polymerases by polymerase chain reaction and kinetic studies on enzymatic production of modified DNA. *Nucleic Acids Res*. 2006;34(19):5383-94.
95. Kuwahara M, Takahata Y, Shoji A, Ozaki AN, Ozaki H, Sawai H. Substrate properties of C5-substituted pyrimidine 2'-deoxynucleoside 5'-triphosphates for thermostable DNA polymerases during PCR. *Bioorg Med Chem Lett*. 2003;13(21):3735-8.
96. Kuwahara M, Takeshima H, Nagashima J, Minezaki S, Ozaki H, Sawai H. Transcription and reverse transcription of artificial nucleic acids involving

backbone modification by template-directed DNA polymerase reactions. *Bioorg Med Chem*. 2009;17(11):3782-8.

97. Lachmann D, Berndl S, Wolfbeis OS, Wagenknecht HA. Synthetic incorporation of Nile Blue into DNA using 2'-deoxyriboside substitutes: Representative comparison of (R)- and (S)-aminopropanediol as an acyclic linker. *Beilstein J Org Chem*. 2010;6:13.
98. Langer PR, Waldrop AA, Ward DC. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proc Natl Acad Sci U S A*. 1981;78(11):6633-7.
99. Lapa SA, Chudinov AV, Timofeev EN. The Toolbox for Modified Aptamers. *Mol Biotechnol*. 2016;58(2):79-92.
100. Latham JA, Johnson R, Toole JJ. The application of a modified nucleotide in aptamer selection: novel thrombin aptamers containing 5-(1-pentynyl)-2'-deoxyuridine. *Nucleic Acids Res*. 1994;22(14):2817-22.
101. Le BH, Koo JC, Joo HN, Seo YJ. Diverse size approach to incorporate and extend highly fluorescent unnatural nucleotides into DNA. *Bioorg Med Chem*. 2017;25(14):3591-3596.
102. Lee LG, Connell CR, Woo SL, Cheng RD, McArdle BF, Fuller CW, Halloran ND, Wilson RK. DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic Acids Res*. 1992;20(10):2471-83.
103. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, Imielinski M; PCAWG Structural Variation Working Group, Weischenfeldt J, Beroukhim R, Campbell PJ; PCAWG Consortium. Patterns of somatic structural variation in human cancer genomes. *Nature*. 2020;578(7793):112-121.
104. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol*. 2009;27(7):652-8.
105. Liu E, Lam CH, Perrin DM. Synthesis and Enzymatic Incorporation of Modified Deoxyuridine Triphosphates. *Molecules*. 2015;20(8):13591-602.
106. Ma F, Fuqua BK, Hasin Y, Yukhtman C, Vulpe CD, Lusis AJ, Pellegrini M. A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC Genomics*. 2019;20(1):9.
107. Mačková M, Pohl R, Hocek M. Polymerase synthesis of DNAs bearing vinyl groups in the major groove and their cleavage by restriction endonucleases. *Chembiochem*. 2014;15(15):2306-12.

108. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214.

109. Maeda M, Shimada T, Ishihama A. Strength and Regulation of Seven rRNA Promoters in *Escherichia coli*. *PLoS One*. 2015;10(12):e0144697.

110. Mäkinen JJ, Shin Y, Vieras E, Virta P, Metsä-Ketelä M, Murakami KS, Belogurov GA. The mechanism of the nucleo-sugar selection by multi-subunit RNA polymerases. *Nat Commun*. 2021;12(1):796.

111. Malla MA, Dubey A, Kumar A, Yadav S, Hashem A, Abd Allah EF. Exploring the Human Microbiome: The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Front Immunol*. 2019;9:2868.

112. Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods*. 2010;7(2):130-2.

113. Manoharan M. 2'-carbohydrate modifications in antisense oligonucleotide therapy: importance of conformation, configuration and conjugation. *Biochim Biophys Acta*. 1999;1489(1):117-30.

114. Marx A, Betz K. The Structural Basis for Processing of Unnatural Base Pairs by DNA Polymerases. *Chemistry*. 2020;26(16):3446-3463.

115. Matray TJ, Kool ET. A specific partner for abasic damage in DNA. *Nature*. 1999;399(6737):704-8.

116. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*. 2010;466(7303):253-7.

117. Maxwell SM, Colls P, Hodes-Wertz B, McCulloh DH, McCaffrey C, Wells D, Munné S, Grifo JA. Why do euploid embryos miscarry? A case-control study comparing the rate of aneuploidy within presumed euploid embryos that resulted in miscarriage or live birth using next-generation sequencing. *Fertil Steril*. 2016;106(6):1414-1419.e5.

118. Meek KN, Rangel AE, Heemstra JM. Enhancing aptamer function and stability via *in vitro* selection using modified nucleic acids. *Methods*. 2016;106:29-36.

119. Menéndez-Arias L. Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. *Viruses*. 2009;1(3):1137-65.

120. Menzel U, Greiff V, Khan TA, Haessler U, Hellmann I, Friedensohn S, Cook SC, Pogson M, Reddy ST. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS One*. 2014;9(5):e96727.

121. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, Batlle E, Sagar, Grün D, Lau JK, Boutet SC, Sanada C, Ooi A, Jones RC, Kaihara K, Brampton C, Talaga Y, Sasagawa Y, Tanaka K, Hayashi T, Braeuning C, Fischer C, Sauer S, Trefzer T, Conrad C, Adiconis X, Nguyen LT, Regev A, Levin JZ, Parekh S, Janjic A, Wange LE, Bagnoli JW, Enard W, Gut M, Sandberg R, Nikaido I, Gut I, Stegle O, Heyn H. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol*. 2020;38(6):747-755.

122. Mikutis S, Gu M, Sendinc E, Hazemi ME, Kiely-Collins H, Aspris D, Vassiliou GS, Shi Y, Tzelepis K, Bernardes G. meCLICK-Seq, a Substrate-Hijacking and RNA Degradation Strategy for the Study of RNA Methylation. *ACS Cent. Sci*. 2020, doi:10.1021/acscentsci.0c01094.

123. Milisavljevič N, Perlíková P, Pohl R, Hocek M. Enzymatic synthesis of base-modified RNA by T7 RNA polymerase. A systematic study and comparison of 5-substituted pyrimidine and 7-substituted 7-deazapurine nucleoside triphosphates as substrates. *Org Biomol Chem*. 2018;16(32):5800-5807.

124. Miller EM, Patterson NE, Zechmeister JM, Bejerano-Sagie M, Delio M, Patel K, Ravi N, Quispe-Tintaya W, Maslov A, Simmons N, Castaldi M, Vijg J, Karabakhtsian RG, Greally JM, Kuo DYS, Montagna C. Development and validation of a targeted next generation DNA sequencing panel outperforming whole exome sequencing for the identification of clinically relevant genetic variants. *Oncotarget*. 2017;8(60):102033-102045.

125. Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D. Iso-Seq Allows Genome-Independent Transcriptome Profiling of Grape Berry Development. *G3 (Bethesda).* 2019;9(3):755-767.

126. Miura F, Fujino T, Kogashi K, Shibata Y, Miura M, Isobe H, Ito T. Triazole linking for preparation of a next-generation sequencing library from single-stranded DNA. *Nucleic Acids Res*. 2018;46(16):e95.

127. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5(7):621-8.

128. Motameny S, Wolters S, Nürnberg P, Schumacher B. Next Generation Sequencing of miRNAs - Strategies, Resources and Methods. *Genes (Basel)*. 2010;1(1):70-84.
129. Motea EA, Berdis AJ. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochim Biophys Acta*. 2010;1804(5):1151-66.
130. Nakatani K., Tor Y. Modified Nucleic Acids, Nucleic Acids and Molecular Biology Series; Springer, 2016; Vol. 31:1-276.
131. Niemeyer CM. Semisynthetic DNA-protein conjugates for biosensing and nanofabrication. *Angew Chem Int Ed Engl*. 2010;49(7):1200-16.
132. Obeid S, Baccaro A, Welte W, Diederichs K, Marx A. Structural basis for the synthesis of nucleobase modified DNA by *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci U S A*. 2010;107(50):21327-31.
133. Ochoa S, Milam VT. Modified Nucleic Acids: Expanding the Capabilities of Functional Oligonucleotides. *Molecules*. 2020;25(20):4659.
134. Østergaard ME, Guenther DC, Kumar P, Baral B, Deobald L, Paszczynski AJ, Sharma PK, Hrdlicka PJ. Pyrene-functionalized triazole-linked 2'-deoxyuridines-probes for discrimination of single nucleotide polymorphisms (SNPs). *Chem Commun (Camb)*. 2010;46(27):4929-31.
135. Ozsolak F, Milos PM. Transcriptome profiling using single-molecule direct RNA sequencing. *Methods Mol Biol*. 2011;733:51-61.
136. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. Direct RNA sequencing. *Nature*. 2009;461(7265):814-8.
137. Panattoni A, Pohl R, Hocek M. Flexible Alkyne-Linked Thymidine Phosphoramidites and Triphosphates for Chemical or Polymerase Synthesis and Fast Postsynthetic DNA Functionalization through Copper-Catalyzed Alkyne-Azide 1,3-Dipolar Cycloaddition. *Org Lett*. 2018;20(13):3962-3965.
138. Pavesi G. ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks. *Adv Biochem Eng Biotechnol*. 2017;160:1-14.
139. Pawar MG, Srivatsan SG. Synthesis, photophysical characterization, and enzymatic incorporation of a microenvironment-sensitive fluorescent uridine analog. *Org Lett*. 2011;13(5):1114-7.
140. Perlíková P, Rylová G, Nauš P, Elbert T, Tloušťová E, Bourderioux A, Slavětínská LP, Motyka K, Doležal D, Znojek P, Nová A, Harvanová M, Džubák P, Šiller M, Hlaváč J, Hajdúch M, Hocek M. 7-(2-Thienyl)-7-Deazaadenosine (AB61), a New Potent Nucleoside Cytostatic with a Complex Mode of Action. *Mol Cancer Ther*. 2016;15(5):922-37.

141. Perrin DM, Garestier T, Hélène C. Expanding the catalytic repertoire of nucleic acid catalysts: simultaneous incorporation of two modified deoxyribonucleoside triphosphates bearing ammonium and imidazolyl functionalities. *Nucleosides Nucleotides*. 1999;18(3):377-91.

142. Pfeiffer F, Tolle F, Rosenthal M, Brändle GM, Ewers J, Mayer G. Identification and characterization of nucleobase-modified aptamers by click-SELEX. *Nat Protoc*. 2018;13(5):1153-1180.

143. Pradhan B, Sarvilinna N, Matilainen J, Aska E, Sjöberg J, Kauppi L. Detection and screening of chromosomal rearrangements in uterine leiomyomas by long-distance inverse PCR. *Genes Chromosomes Cancer*. 2016;55(3):215-26.

144. Precone V, Del Monaco V, Esposito MV, De Palma FD, Ruocco A, Salvatore F, D'Argenio V. Cracking the Code of Human Diseases Using Next-Generation Sequencing: Applications, Challenges, and Perspectives. *Biomed Res Int.* 2015;2015:161648.

145. Pu W, Wang C, Chen S, Zhao D, Zhou Y, Ma Y, Wang Y, Li C, Huang Z, Jin L, Guo S, Wang J, Wang M. Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC). *Clin Epigenetics*. 2017;9:129.

146. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41:D590-6.

147. Ramsay N, Jemth AS, Brown A, Crampton N, Dear P, Holliger P. CyDNA: synthesis and replication of highly Cy-dye substituted DNA by an evolved polymerase. *J Am Chem Soc*. 2010;132(14):5096-104.

148. Robertson DL, Joyce GF. Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*. 1990;344(6265):467-8.

149. Rong M, He B, McAllister WT, Durbin RK. Promoter specificity determinants of T7 RNA polymerase. *Proc Natl Acad Sci U S A*. 1998;95(2):515-9.

150. Rosemeyer H, Seela F. Modified purine nucleosides as dangling ends of DNA duplexes: the effect of the nucleobase polarizability on stacking interactions. *J. Chem. Soc., Perkin Trans*. 2002;2:746-750.

151. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, Graybuck LT, Peeler DJ, Mukherjee S, Chen W, Pun SH, Sellers DL, Tasic B, Seelig G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. 2018;360(6385):176-182.

152. Rosenblum BB, Lee LG, Spurgeon SL, Khan SH, Menchen SM, Heiner CR, Chen SM. New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res.* 1997;25(22):4500-4.

153. Routh A, Head SR, Ordoukhanian P, Johnson JE. ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3'-Azido cDNAs. *J Mol Biol*. 2015;427(16):2610-6.

154. Routh A, Ji P, Jaworski E, Xia Z, Li W, Wagner EJ. Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res*. 2017;45(12):e112.

155. Sabat AJ, van Zanten E, Akkerboom V, Wisselink G, van Slochteren K, de Boer RF, Hendrix R, Friedrich AW, Rossen JWA, Kooistra-Smid AMDM. Targeted next-generation sequencing of the 16S-23S rRNA region for culture-independent bacterial identification - increased discrimination of closely related species. *Sci Rep*. 2017;7(1):3434.

156. Sandhu C, Qureshi A, Emili A. Panomics for Precision Medicine. *Trends Mol Med.* 2018;24(1):85-101.

157. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-7.

158. Sanzone AP, El-Sagheer AH, Brown T, Tavassoli A. Assessing the biocompatibility of click-linked DNA in *Escherichia coli. Nucleic Acids Res*. 2012;40(20):10567-75.

159. Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, Kurisaki A, Nikaido I. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol*. 2018;19(1):29.

160. Sasaki N, Izawa M, Watahiki M, Ozawa K, Tanaka T, Yoneda Y, Matsuura S, Carninci P, Muramatsu M, Okazaki Y, Hayashizaki Y. Transcriptional sequencing: A method for DNA sequencing using RNA polymerase. *Proc Natl Acad Sci U S A*. 1998;95(7):3455-60.

161. Sawai H, Ozaki AN, Satoh F, Ohbayashi T, Masud MM, Ozaki H. Expansion of structural and functional diversities of DNA using new 5-substituted deoxyuridine derivatives by PCR with superthermophilic KOD Dash DNA polymerase. *Chem. Commun*. 2001, 2604–2605.

162. Schultz HJ, Gochi AM, Chia HE, Ogonowsky AL, Chiang S, Filipovic N, Weiden AG, Hadley EE, Gabriel SE, Leconte AM. Taq DNA Polymerase Mutants and 2'-Modified Sugar Recognition. *Biochemistry*. 2015;54(38):5999-6008.

163. Sebastián-Martín A, Barrioluengo V, Menéndez-Arias L. Transcriptional inaccuracy threshold attenuates differences in RNA-dependent DNA

synthesis fidelity between retroviral reverse transcriptases. *Sci Rep*. 2018;8(1):627.

164. Seela F, Becher G. Pyrazolo[3,4-d]pyrimidine nucleic acids: adjustment of dA-dT to dG-dC base pair stability. *Nucleic Acids Res*. 2001;29(10):2069-78.
165. Seela F, Pujari SS. Azide-alkyne "click" conjugation of 8-aza-7-deazaadenine-DNA: synthesis, duplex stability, and fluorogenic dye labeling. *Bioconjug Chem*. 2010;21(9):1629-41.
166. Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, Prado JL, Prieto A, Willerslev E, Orlando L. Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLoS One.* 2013;8(10):e78575.
167. Selmi B, Boretto J, Sarfati SR, Guerreiro C, Canard B. Mechanism-based suppression of dideoxynucleotide resistance by K65R human immunodeficiency virus reverse transcriptase using an alpha-boranophosphate nucleoside analogue. *J Biol Chem*. 2001;276(51):48466-72.
168. Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res*. 2010;38(2):391-9.
169. Seyfried P, Heinz M, Pintér G, Klötzner DP, Becker Y, Bolte M, Jonker HRA, Stelzl LS, Hummer G, Schwalbe H, Heckel A. Optimal Destabilization of DNA Double Strands by Single-Nucleobase Caging. *Chemistry*. 2018;24(66):17568-17576.
170. Shaughnessy KH, DeVasher RB. Palladium-Catalyzed Cross-Coupling in Aqueous Media: Recent Progress and Current Applications. *Curr Org Chem.* 2005;9:585-604.
171. Shivalingam A, Tyburn AE, El-Sagheer AH, Brown T. Molecular Requirements of High-Fidelity Replication-Competent DNA Backbones for Orthogonal Chemical Ligation. *J Am Chem Soc.* 2017;139(4):1575-1583.
172. Shoji A, Kuwahara M, Ozaki H, Sawai H. Modified DNA aptamer that binds the (R)-isomer of a thalidomide derivative with high enantioselectivity. *J Am Chem Soc*. 2007;129(5):1456-64.
173. Sinkeldam RW, Greco NJ, Tor Y. Fluorescent analogs of biomolecular building blocks: design, properties, and applications. *Chem Rev*. 2010;110(5):2579-619.
174. Smith CC, Hollenstein M, Leumann CJ. The synthesis and application of a diazirine-modified uridine analogue for investigating RNA–protein interactions. *RSC Adv*. 2014;4:48228-48235.

175. Sonogashira K, Tohda Y, Hagihara N. A convenient synthesis of acetylenes: catalytic substitutions of acetylenic hydrogen with bromoalkenes, iodoarenes and bromopyridines. *Tetrahedron Letters*. 1975;16(50):4467-70.

176. Sørensen RS, Okholm AH, Schaffert D, Kodal AL, Gothelf KV, Kjems J. Enzymatic ligation of large biomolecules to DNA. *ACS Nano*. 2013;7(9):8098-104.

177. Soriano-Lerma A, Pérez-Carrasco V, Sánchez-Marañón M, Ortiz-González M, Sánchez-Martín V, Gijón J, Navarro-Mari JM, García-Salcedo JA, Soriano M. Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Sci Rep*. 2020;10(1):13637.

178. Sousa R, Padilla R. A mutant T7 RNA polymerase as a DNA polymerase. *EMBO J*. 1995;14(18):4609-21.

179. Srivatsan SG, Tor Y. Enzymatic incorporation of emissive pyrimidine ribonucleotides. *Chem Asian J*. 2009;4(3):419-27.

180. Srivatsan SG, Tor Y. Fluorescent pyrimidine ribonucleotide: synthesis, enzymatic incorporation, and utilization. *J Am Chem Soc*. 2007;129(7):2044-53.

181. Stangl C, de Blank S, Renkens I, Westera L, Verbeek T, Valle-Inclan JE, González RC, Henssen AG, van Roosmalen MJ, Stam RW, Voest EE, Kloosterman WP, van Haaften G, Monroe GR. Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nat Commun*. 2020;11(1):2861.

182. Staševskij Z, Gibas P, Gordevičius J, Kriukienė E, Klimašauskas S. Tethered Oligonucleotide-Primed Sequencing, TOP-Seq: A High-Resolution Economical Approach for DNA Epigenome Profiling. *Mol Cell*. 2017;65(3):554-564.e6.

183. Tabor S, Richardson CC. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc Natl Acad Sci U S A*. 1995;92(14):6339-43.

184. Tang DT, Plessy C, Salimullah M, Suzuki AM, Calligaris R, Gustincich S, Carninci P. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Res*. 2013;41(3):e44.

185. Tanpure AA, Srivatsan SG. A microenvironment-sensitive fluorescent pyrimidine ribonucleoside analogue: synthesis, enzymatic incorporation, and fluorescence detection of a DNA abasic site. *Chemistry*. 2011;17(45):12820-7.

186. Tanpure AA, Srivatsan SG. Synthesis, photophysical properties and incorporation of a highly emissive and environment-sensitive uridine analogue based on the Lucifer chromophore. *Chembiochem*. 2014;15(9):1309-16.

187. Tarashima N, Ando H, Kojima T, Kinjo N, Hashimoto Y, Furukawa K, Ishida T, Minakawa N. Gene Silencing Using 4'-thioDNA as an Artificial Template to Synthesize Short Hairpin RNA Without Inducing a Detectable Innate Immune Response. *Mol Ther Nucleic Acids*. 2016;5(1):e274.

188. Temiakov D, Patlan V, Anikin M, McAllister WT, Yokoyama S, Vassylyev DG. Structural basis for substrate selection by T7 RNA polymerase. *Cell*. 2004;116(3):381-91.

189. Thoresen LH, Jiao GS, Haaland WC, Metzker ML, Burgess K. Rigid, conjugated, fluoresceinated thymidine triphosphates: syntheses and polymerase mediated incorporation into DNA analogues. *Chemistry*. 2003;9(19):4603-10.

190. Thorsen J, Micci F, Heim S. Identification of chromosomal breakpoints of cancer-specific translocations by rolling circle amplification and long-distance inverse PCR. *Cancer Genet*. 2011;204(8):458-61.

191. Tolle F, Brändle GM, Matzner D, Mayer G. A Versatile Approach Towards Nucleobase-Modified Aptamers. *Angew Chem Int Ed Engl*. 2015;54(37):10971-4.

192. Troll CJ, Kapp J, Rao V, Harkins KM, Cole C, Naughton C, Morgan JM, Shapiro B, Green RE. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics*. 2019;20(1):1023.

193. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*. 1990;249(4968):505-10.

194. UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006. Accessed 8 Dec. 2020.

195. Vaish NK, Fraley AW, Szostak JW, McLaughlin LW. Expanding the structural and functional diversity of RNA: analog uridine triphosphates as candidates for in vitro selection of nucleic acids. *Nucleic Acids Res*. 2000;28(17):3316-22.

196. Vaught JD, Dewey T, Eaton BE. T7 RNA polymerase transcription with 5-position modified UTP derivatives. *J Am Chem Soc*. 2004;126(36):11231-7.

197. Velten L, Anders S, Pekowska A, Järvelin AI, Huber W, Pelechano V, Steinmetz LM. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol Syst Biol*. 2015;11(6):812.

198. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 2004;304(5667):66-74.

199. Verga D, Welter M, Marx A. Sequence selective naked-eye detection of DNA harnessing extension of oligonucleotide-modified nucleotides. *Bioorg Med Chem Lett*. 2016;26(3):841-844.

200. Verga D, Welter M, Steck AL, Marx A. DNA polymerase-catalyzed incorporation of nucleotides modified with a G-quadruplex-derived DNAzyme. *Chem Commun (Camb)*. 2015;51(34):7379-81.

201. Walunj MB, Tanpure AA, Srivatsan SG. Post-transcriptional labeling by using Suzuki-Miyaura cross-coupling generates functional RNA probes. *Nucleic Acids Res*. 2018;46(11):e65.

202. Weber S, Büscher AK, Hagmann H, Liebau MC, Heberle C, Ludwig M, Rath S, Alberer M, Beissert A, Zenker M, Hoyer PF, Konrad M, Klein HG, Hoefele J. Dealing with the incidental finding of secondary variants by the example of SRNS patients undergoing targeted next-generation sequencing. *Pediatr Nephrol*. 2016;31(1):73-81.

203. Welter M, Verga D, Marx A. Sequence-Specific Incorporation of Enzyme-Nucleotide Chimera by DNA Polymerases. *Angew Chem Int Ed Engl.* 2016;55(34):10131-5.

204. Whitfield CJ, Little RC, Khan K, Ijiro K, Connolly BA, Tuite EM, Pike AR. Self-Priming Enzymatic Fabrication of Multiply Modified DNA. *Chemistry*. 2018;24(57):15267-15274.

205. Widschwendter M, Evans I, Jones A, Ghazali S, Reisel D, Ryan A, Gentry-Maharaj A, Zikan M, Cibula D, Eichner J, Alunni-Fabbroni M, Koch J, Janni WJ, Paprotka T, Wittenberger T, Menon U, Wahl B, Rack B, Lempiäinen H. Methylation patterns in serum DNA for early identification of disseminated breast cancer. *Genome Med*. 2017;9(1):115.

206. Winand R, Bogaerts B, Hoffman S, Lefevre L, Delvoye M, Braekel JV, Fu Q, Roosens NH, Keersmaecker SC, Vanneste K. Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and third (Oxford Nanopore Technologies) generation sequencing technologies. *Int J Mol Sci*. 2019;21(1):298.

207. Winz ML, Samanta A, Benzinger D, Jäschke A. Site-specific terminal and internal labeling of RNA by poly(A) polymerase tailing and copper-catalyzed or copper-free strain-promoted click chemistry. *Nucleic Acids Res.* 2012;40(10):e78.

208. Wolk SK, Shoemaker RK, Mayfield WS, Mestdagh AL, Janjic N. Influence of 5-N-carboxamide modifications on the thermodynamic stability of oligonucleotides. *Nucleic Acids Res*. 2015;43(19):9107-22.

209. Wynne SA, Pinheiro VB, Holliger P, Leslie AG. Structures of an apo and a binary complex of an evolved archeal B family DNA polymerase capable of synthesising highly cy-dye labelled DNA. *PLoS One*. 2013;8(8):e70892.

210. Xiong Y, Soumillon M, Wu J, Hansen J, Hu B, van Hasselt JGC, Jayaraman G, Lim R, Bouhaddou M, Ornelas L, Bochicchio J, Lenaeus L, Stocksdale J, Shim J, Gomez E, Sareen D, Svendsen C, Thompson LM, Mahajan M, Iyengar R, Sobie EA, Azeloglu EU, Birtwistle MR. A Comparison of mRNA Sequencing with Random Primed and 3'-Directed Libraries. *Sci Rep.* 2017;7(1):14626.

211. Xu H, Fair BJ, Dwyer ZW, Gildea M, Pleiss JA. Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nat Methods*. 2019;16(1):55-58.

212. Xu W, Chan KM, Kool ET. Fluorescent nucleobases as tools for studying DNA and RNA. *Nat Chem*. 2017;9(11):1043-1055.

213. Zhang Y, Ji P, Wang J, Zhao F. RiboFR-Seq: a novel approach to linking 16S rRNA amplicon profiles to metagenomes. *Nucleic Acids Res*. 2016;44(10):e99.

214. Zhao C, Liu F, Pyle AM. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA*. 2018;24(2):183-195.

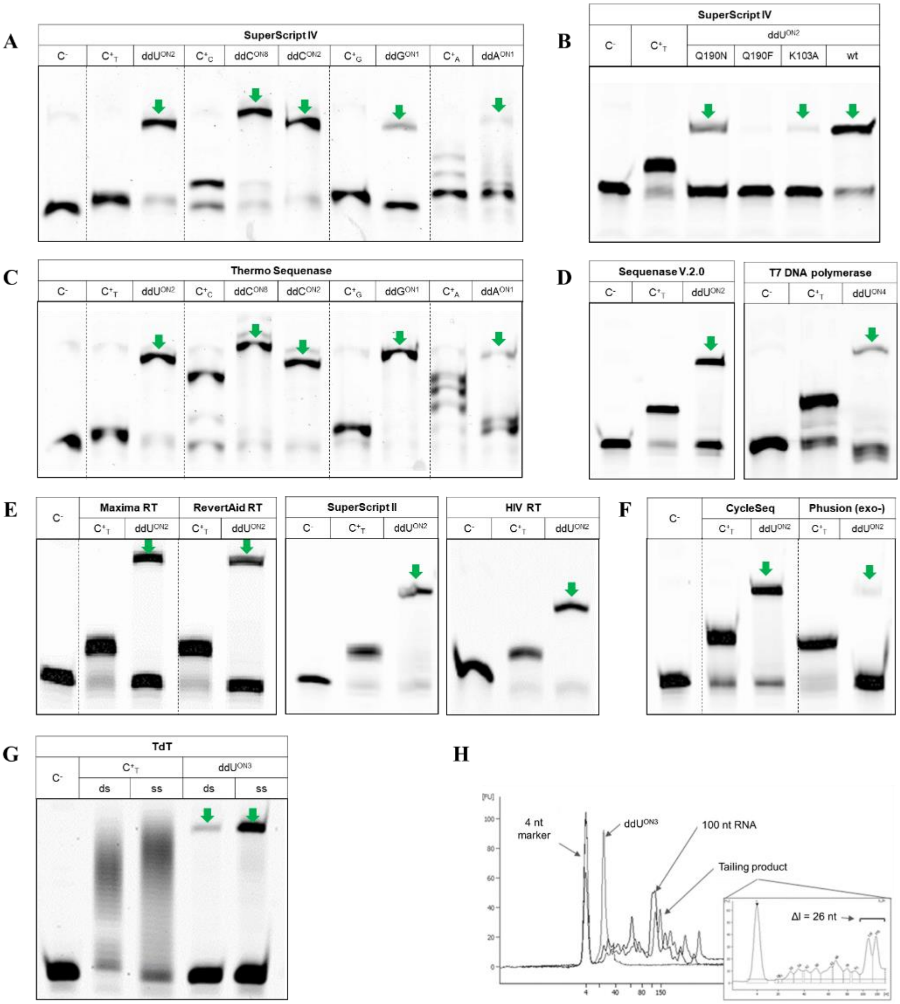
215. Zheng Y, Beal PA. Synthesis and evaluation of an alkyne-modified ATP analog for enzymatic incorporation into RNA. *Bioorg Med Chem Lett*. 2016;26(7):1799-802.

216. Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, Chen J, Robinson HE, Shim HS, Chmielecki J, Pao W, Engelman JA, Iafrate AJ, Le LP. Anchored multiplex PCR for targeted next-generation sequencing. *Nat Med*. 2014;20(12):1479-84.

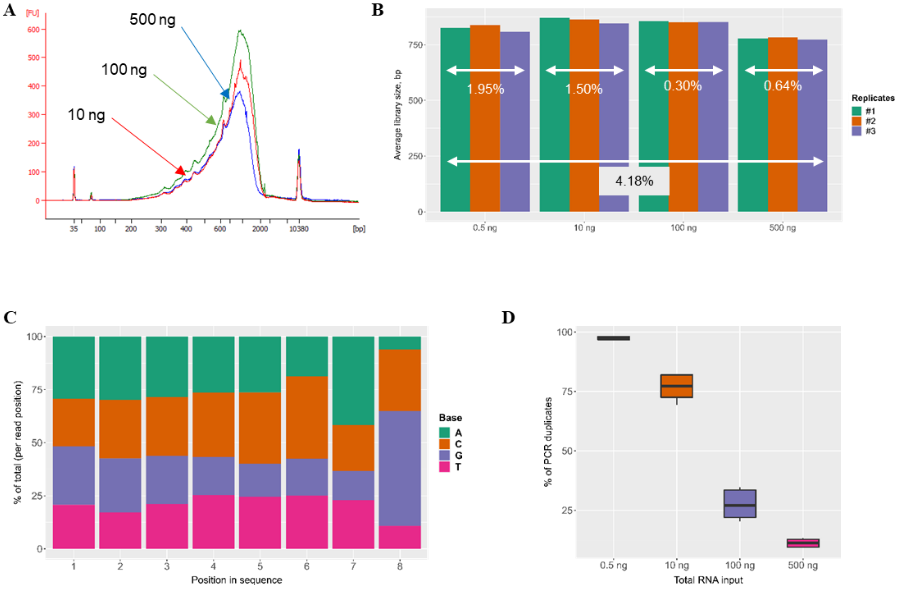
217. Zhou J, Rossi J. Aptamers as targeted therapeutics: current potential and challenges. *Nat Rev Drug Discov.* 2017;16(3):181-202.

218. Zhu B, Hernandez A, Tan M, Wollenhaupt J, Tabor S, Richardson CC. Synthesis of 2'-Fluoro RNA by Syn5 RNA polymerase. *Nucleic Acids Res*. 2015;43(14):e94.

219. Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res*. 2012;40(7):e54.



Supplementary Figure 1. Incorporation of OTDDNs by DNA and RNA polymerases. **(A)** – incorporation by SuperScript IV RT. **(B)** – incorporation by SuperScript IV mutant variants. **(C)** – incorporation by Thermo Sequenase. **(D)** – incorporation by wild type T7 DNA polymerase and its engineered variant Sequenase V2.0. **(E)** – incorporation by various reverse transcriptases. **(F)** – incorporation by thermostable DNA polymerases CycleSeq and Phusion (exo-). **(G)** – incorporation by TdT into double-stranded DNA duplex (ds) or single-stranded ON (ss). **(H)** – incorporation by PUP RNA polymerase. C^- - negative control; $C^+_{T/C/G/A}$ – positive control.



Supplementary Figure 2. The robustness of OTDDN incorporation rate across various RNA inputs and the use of a randomized region within OTDDN as a UMI. **(A)** – MTAS-seq generates libraries of very similar traces from various amounts of total RNA. **(B)** – the coefficients of variation of the average library size across different RNA inputs and technical replicates. **(C)** – typical base composition within UMI region. **(D)** – fraction of PCR duplicates identified by UMIs across genes for a series of MTAS-seq libraries prepared from different amounts of starting material.

NOTES

NOTES

NOTES

Vilnius University Press
9 Saulėtekio Ave., Building III, LT-10222 Vilnius
Email: info@leidykla.vu.lt, www.leidykla.vu.lt
Print run 15 copies