BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

MASTER'S THESIS

Brno, 2020 Bc. Lujza Barilíková

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

PROCESSING OF UNIQUE MOLECULAR IDENTIFIERS WITHOUT MAPPING TO A REFERENCE GENOME
ZPRACOVÁNÍ UNIKÁTNÍCH MOLEKULÁRNÍCH INDEXŮ BEZ MAPOVÁNÍ K REFERENČNÍMU GENOMU

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR — Bc. Lujza Barilíková
AUTOR PRÁCE

SUPERVISOR — Mgr. Ing. Karel Sedlář, Ph.D.
VEDOUCÍ PRÁCE

BRNO 2020

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Master's Thesis

Master's study program **Biomedical Engineering and Bioinformatics**

Department of Biomedical Engineering

Student: Bc. Lujza Barilíková ***ID:*** 185945

Year of study: 2 ***Academic year:*** 2019/20

TITLE OF THESIS:

Processing of Unique Molecular Identifiers without Mapping to a Reference Genome

INSTRUCTION:

1) Prepare a literature review of bioinformatics processing techniques for recognition of unique molecular identifiers (UMI) in next generation sequencing data. 2) Study the possibility of UMI inexact clustering and creation of consensus sequences from particular clusters. 3) Propose a reference-free UMI processing method for removing PCR duplicates. 4) Implement the method in a selected programming language. 5) Compare your method to existing methods, evaluate speed and accuracy of different approaches. 6) Discuss the results.

RECOMMENDED LITERATURE:

[1] ISLAM, Saiful, Amit ZEISEL, Simon JOOST, Gioele LA MANNO, Pawel ZAJAC, Maria KASPER, Peter LÖNNERBERG a Sten LINNARSSON. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods, 2014, 11(2), 163-166.

[2] SMITH, Tom, Andreas HEGER a Ian SUDBERY. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research, 2017, 27(3), 491-499.

Date of project specification: 3.2.2020 ***Deadline for submission:*** 29.5.2020

Supervisor: Mgr. Ing. Karel Sedlář, Ph.D.

prof. Ing. Ivo Provazník, Ph.D.
Chair of study program board

ABSTRACT

The main purpose of this thesis is to design a new algorithm for processing unique molecular identifiers (UMIs) without mapping to a reference genome. These random oligonucleotide sequences are attracting an increasing interest due to its ability to facilitate PCR error and bias recognition. Since there has been a rapid rise in the use of next-generation sequencing (NGS) technologies, great effort has been put into the development of tools for data analysis. At present, tools to solve these errors are usually relative time-consuming and complex due to computationally demanding alignment. The most important limitation of these tools lies in the fact that multi-mapping reads are allowed when processing duplicates. These reads are usually ignored and may lead to reduction of quantitative accuracy and cause misleading interpretation of sequencing results. In order to solve this problem, a new approach is introduced in this thesis, which allows estimating the absolute number of unique molecules with relatively fast and reliable performance.

KEYWORDS

unique molecular identifier (UMI), next-generation sequencing, PCR error, duplicates

ABSTRAKT

Hlavným cieľom tejto práce je návrh nového algoritmu k spracovaniu unikátnych molekulárnych indexov bez mapovania na referenčný genóm. O tieto náhodné oligonukleotidové sekvencie neustále vzrastá záujem, pretože uľahčujú rozpoznávať PCR chyby a skresľovanie údajov. Keďže používanie technológií sekvenovania novej generácie neustále rastie, je vynaložené veľké úsilie vyvíjať nástroje pre analýzu produkovaných dát. V súčasnosti sú nástroje na riešenie týchto chýb relatívne časovo náročné a zložité z dôvodu výpočtovo náročného zarovnania. Najdôležitejšie obmedzenie týchto nástrojov spočíva v skutočnosti, že pri spracovávaní duplikátov sú povolené multi-mapované čítania. Tieto čítania sú zvyčajne ignorované, čo môže viesť k zníženiu kvantitatívnej presnosti a spôsobiť zavádzajúcu interpretáciu výsledkov daného sekvenovania. V snahe vyriešiť tento problém je v tejto práci uvedený nový prístup, ktorý umožňuje odhad absolútneho počtu jedinečných molekúl s relatívne rýchlym a spoľahlivým spôsobom.

KLÍČOVÁ SLOVA

unikátne molekulárne identifikátory (UMI), nová generácia sekvenovania, PCR chyby, duplikáty

BARILÍKOVÁ, Lujza. *Processing of Unique Molecular Identifiers without Mapping to a Reference Genome*. Brno, 2020. Available from: `https://www.vutbr.cz/studenti/zav-prace/detail/126827`. Master's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering. Advised by Mgr. Ing. Karel Sedlář, Ph.D.

ROZŠÍRENÝ ABSTRAKT

Je známe, že sekvenovanie DNA alebo RNA sa stáva neustále dôležitejším a vplyvnejším ako kedykoľvek predtým. V posledných rokoch prudko narástlo využívanie technológií sekvenovania, s čím súvisí aj pokrok nástrojov vyžívaných v bioinformatike. V priebehu nasledujúcich niekoľkých rokov sa vysokovýkonné sekvenovanie pravdepodobne stane neoddeliteľnou súčasťou genomického, epigenomického, proteomického, transkriptomického a metabolomického výskumu, pretože umožňuje náhľad k molekulárnemu mechanizmu pôsobiaceho na reguláciu genómu. Nedávne udalosti kvantitatívnej analýzy sekvenačných dát, produkovaných technológiami novej generácie, naznačujú isté znepokojenie zo značného skreslenia a chyby spôsobenej PCR amplifikáciou vybraných regiónov počas prípravy knižnice. Predchádzajúce nástroje využívané v bioinformatike sa pri riešení tohto problému obmedzili iba na sekvenačnú identitu a lokalitu mapovania sekvencií, s cieľom znížiť počet duplikátov PCR. Tento prístup je však pomerne zjednodušujúci a ignoruje niektoré biologické aspekty spracovávaných dát a môže viesť k nepresným záverom. Na vyriešenie tohto problému sa v procese sekvenovania začali využívať začleňujúce sa náhodné sekvencie, známe ako unikátne molekulárne identifikátory (UMI). Požitie týchto unikátnych molekulárnych identifikátorov sa postupne začalo využívať v mnohých aplikáciách, pretože poskytujú mnoho výhod. Napriek prítomnosti niekoľkých algoritmov, neustále existuje potreba inovácie účinnejších nástrojov, pomocou ktorých by bolo možné odhadnúť absolútny počet jedinečných molekúl. Cieľom tejto práce je preto navrhnúť algoritmus k spracovaniu UMI, kombinujúci dostatočný výkon s relatívnou jednoduchosťou.

Úvodná časť tejto práce poskytuje stručný prehľad prístupov sekvenovania DNA a ich aplikácií. Vo všeobecnosti je sekvenovanie DNA metóda určovania poradia nukleových báz v molekule DNA. Každý jednotlivec a organizmus má špecifickú nukleotidovú sekvenciu a v súvislosti s tým môže DNA sekvenovanie poskytnúť pohľad na rozmanitosť a vývoj organizmov, ktoré nie je možné kultivovať v laboratóriu a tým je znemožnené ich touto cestou študovať. Analýzou genómov sú identifikované gény a regulačné prvky spolu s porozumením ich úloh vo vývoji a evolúcii. Ďalšia časť práce skúma generácie sekvenačných technológií, ktoré umožňujú simultánnu analýzu veľkého počtu sekvencií. S dostupnosťou týchto technológií je možné študovať a analyzovať štruktúru nukleových kyselín pre konkrétne aplikácie, či už v oblasti vedy, klinickej diagnostiky alebo metagenomiky. V posledných rokoch sa tempo pokroku zvýšilo a boli vyvinuté nové technológie, vedúce k exponenciálnemu poklesu cien za súčasného zrýchlenia sekvenačného procesu. Nasledujúca časť práce popisuje kľúčové využitie unikátnych molekulárnych identifikátorov v súvislosti s RNA sekvenovaním v snahe odlíšiť technickú a biologickú duplikáciu analyzovaných molekúl.

V posledných rokoch sa totiž RNA sekvenovanie, využívajúce predovšetkým sekvenačné platformy novej generácie, stalo vysoko používanou metódou analýzy celého transkriptómu. Nielen metódy sekvenovania RNA, ale mnoho ďalších výkonných sekvenačných platforiem vyžaduje, aby sa v priebehu prípravy knižnice pred samotným sekvenovaním vykonala amplifikácia analyzovaných molekúl formou PCR. Avšak, všetky molekuly sú amplifikované s odlišnou pravdepodobnosťou, čo vo výsledku môže viesť k tomu, že niektoré molekuly sú v pripravenej knižnici prezentované v nadmernej miere v porovnaní s ostatnými molekulami. Na rozlíšenie medzi identickými kópiami pochádzajúcimi z odlišných molekúl a duplikátmi PCR pochádzajúcimi z tej istej molekuly sa používajú spomínané krátke náhodné oligonukleotidové sekvencie, teda UMI. Nástroje spracovávajúce UMI k dosiahnutiu deduplikácie sekvencií, vo všeobecnosti začínajú proces spracovania časovo náročným mapovaním sekvencií k referenčnému genómu. Navyše sú v priebehu zarovnávania typicky povolené viacnásobné mapovania, ktoré sú definované ako sekvencie, mapujúce sa rôzne miesta genómu v dôsledku viacerých kópií génu. To sťažuje rozlíšenie medzi skutočne viacnásobným mapovaním a čítaniami, ktoré pochádzajú z viacerých fragmentov toho istého génu. Mnoho nástrojov tieto sekvencie typicky ignoruje, čo znamená, že najmenej 20 - 30 % dát je zanedbaných. Dizajn prezentovanej metódy My_UMI_tool je založený na komplexnej štúdii výhod každého vybraného dostupného nástroja. Dosiahnuté poznatky o nástrojoch poskytujú aj informácie o ich nevýhodách a dôvodoch, prečo pretrváva záujem neustále inovovať a vyvíjať nové nástroje. Navrhovaná metóda je implementovaná v programovacom prostredí R pozostáva z nasledujúcich krokov: predbežné spracovanie dát zo vstupného súboru vo formáte FASTQ, zhlukovanie sekvencií s rovnakou UMI do klastrov, následné zoskupovanie sekvencií z klastrov s rovnakými UMI podľa ich podobnosti, určenie počtu počiatočných nezhôd zo zarovnania týchto sekvencií, oprava chýb UMI a konečná identifikácia duplikátov k vygenerovaniu konečného súboru FASTQ s deduplikovanými čítaniami a TSV súbor obsahujúcimi všetky čítania, z ktorých každé má priradenú skupinu, ktorá mu bola v priebehu spracovania pridelená. Jedným z hlavných problémov v rámci chýb vyskytujúcich sa v UMI, ktoré sú výsledkom nukleotidových substitúcií počas PCR, prípadne nukleotidových inzercií alebo delécií počas sekvenovania, je to, že vznikajú falošné UMI, čo môže mať negatívny vplyv na odhad počtu jedinečných molekúl. Aby sa znížila pravdepodobnosť nesprávneho priradenia klastrov jednotlivých sekvencií a zlepšila sa kvantifikácia pomocou UMI, chyby vyskytujúce sa v UMI nie sú ignorované. V snahe vyhodnotiť výkon a efektivitu metódy My_UMI_tool, bol vyššie popísaný algoritmus testovaný na šiestich simulovaných genómických, ako aj dvoch experimentálnych dátach. Účelom simulácie je vygenerovať syntetické genómické dáta, ktorých pôvodný UMI je známy.

Navrhovanú simuláciu je možné považovať za dvojfázový proces, pozostávajúci z generovania biologickej duplikácie a generovania technickej duplikácie. Najprv sa cieľové sekvencie z požadovaného vstupného referenčného súboru FASTA fragmentujú na požadovanú dĺžku s využitím posuvného okna s veľkosťou 75 nukleotidov a s veľkosťou kroku 1 nukleotid. S cieľom simulovania biologickej duplikácia sa získané sekvencie náhodne replikujú. V ďalšom kroku je ku každej jednotlivej sekvencii pripojená náhodná sekvencia UMI. Technická duplikácia je uskutočnená využitím simulátora produkujúceho čítania novej generácie sekvenovania, napodobňovaním skutočného procesu sekvenovania zahŕňajúc chyby, ktoré v tomto procese vznikajú. Metóda je porovnávaná s nástrojom UMI-tools, ktorý patrí medzi najbežnejšie používané nástroje v oblasti spracovania UMI, v rámci ktorej poskytuje predikciu s vysokou presnosťou. K porovnávaniu výkonu nástroja My_UMI_tool bol implementovaný automatický porovnávací postup. Tento postup zahŕňa zvolenie vstupných FASTQ súborov, ktoré majú byť spracované. Tieto súbory sú následné spracované dvoma rôznymi spôsobmi k získaniu dvoch rôznych výsledkov poskytnutých dvoma rôznymi nástrojmi, a to navrhovaným My_UMI_tool a porovnávacím UMI-tools. V prípade UMI-tools sú dáta prvotne zarovnané a následne spracované týmto nástrojom k poskytnutiu výsledných deduplikovaných dát. V prípade My_UMI_tool sú dáta najprv deduplikované a až následne zarovnané príslušným zarovnávacím nástrojom.

Na základe tejto práce je možné vyvodiť záver, že My_UMI_tool je nezanedbateľným nástrojom na deduplikáciu sekvenačných dát novej generácie využívajúcich UMI, z ktorých sú duplikátne čítania zo vzorky odstránené s cieľom pripraviť tieto dáta k následnej analýze. Na rozdiel od existujúcich nástrojov je My_UMI_tool navrhnutý tak, aby sa predišlo mapovaniu sekvencií pred samotnou deduplikáciou, čím sa stáva jedinečným v ponuke momentálne dostupných nástrojov. Výsledky naznačujú, že vynechanie časovo náročného mapovania sekvencií pred deduplikáciou nemá vplyv na konečné stanovenie absolútneho počtu jedinečných molekúl a konečné výsledky sú rovnaké alebo lepšie ako výsledky, ktoré sú v súčasnosti akceptované nástrojom, ktorý bol k porovnávaniu využitý. Napriek tomu, že časová výkonnosť nie je ideálna, sa predpokladá, že nástroj bude užitočný v aplikáciách, ako je napríklad analýza transpozibilných elementov alebo elementov Alu, ktoré tvoria viac ako 10 % ľudského genómu. Z tohto hľadiska je získanie súboru správne duplikovaných čítaní pred zarovnaním rozhodujúce, čo významne rieši problém spracovania viacnásobne mapovaných čítaní.

DECLARATION

I declare that I have written the Master's thesis titled "Processing of Unique Molecular Identifiers without Mapping to a Reference Genome" independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author I furthermore declare that, with respect to the creation of this Master's thesis, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno ....

author's signature

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my advisor Mgr. Ing. Karel Sedlář, Ph.D. and for the continuous support of my Master's thesis, for his patience, motivation and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I would also like to express my eternal appreciation towards my laboratory colleagues from Bioinformatics Core Facility - CEITEC MU. This thesis becomes a reality with their kind support, help and time spent on brainstorming with me.

Brno ....

author's signature

Contents

List of Figures

List of Tables

Introduction

It is well known that DNA or RNA sequencing has become more and more important and influential than ever before. For the past few years, there has been a rapid rise in the use of sequencing technologies and the progress in bioinformatics tools. Within the next few years, high-throughput sequencing is likely to become an inseparable component of genomic, epigenomic, proteomic, transcriptomic as well as metabolomic research as it gives access to a precise picture of the molecular mechanism acting upon genome regulation.

In the light of recent events in quantitative analysis of next-generation sequencing data, there is considerable concern about bias and error introduced by PCR amplification of the targets of interest during library preparation. Previous bioinformatics pipelines have only been limited to sequence identity and sequence alignment to reduce the number of PCR duplicates, which is quite over-simplistic, and ignores some biological aspects of the data and may lead to biased conclusions. To solve this issue, random sequences, known as unique molecular identifiers (UMIs), are incorporated into sequencing workflows. Quite recently, the use of these unique molecular identifiers have been utilized in many applications and provides many benefits, however, there is still a need for more efficient counting algorithm with which one can estimate the absolute number of unique molecules in large input data sets. Therefore, the aim of this work is to design an algorithm which combines a sufficient performance with low complexity.

This thesis is divided into seven sections. The first section gives a brief overview of DNA sequencing approaches and its applications. The second section examines generations of sequencing technologies, which allows simultaneous analysis of a large number of sequences. This is followed by the third section, which explains use of UMIs in the field of RNA-sequencing. In the fourth section, a case study of unique molecular identifiers and the most used tools for dealing with UMIs in next-generation sequencing data sets is presented. Afterwards, a new methodology for handling mentioned UMIs is outlined in the fifth section. Data simulation and obtained final results, together with the evaluation of the performance of the proposed method are discussed in the sixth section and are followed with the conclusions drawn in the final section of the thesis.

1 DNA sequencing and its applications

It is well known that DNA sequencing is a method to determine the order of the four nucleotide bases (adenine, guanine, cytosine and thymine) in an oligonucleotide molecule. As each individual and organism has a specific nucleotide sequence, DNA sequencing can provide insights into diversity and evolution of organisms that cannot be grown in cultures in the laboratory and therefore are not easy to study [1]. The completion of a human genome reference sequence allowed for the development of many genome sequencing instruments [2]. By analyzing genomes, identification of genes and regulatory elements together with the understanding of their roles in development and evolution is obtained. It is also hoped that growing knowledge of the human genome will provide the health tendencies or disease risks of each individual.

1.1 Whole-genome sequencing

Whole-genome sequencing is a comprehensive method that enables to examine the entire genomic DNA sequence of a cell at a single time including coding, non-coding regions, and mtDNA. On the other hand, whole-exome sequencing offers regional genomic sequencing, but only targeted view of the protein-coding regions is acquired [3]. It is expected that through identification of regions of the genome and genetic variants, which are potentially responsible for human evolution, genetic diversity as well as for various diseases, may improve medical diagnostics [4]. Unlike targeted sequencing, whole-genome sequencing provides base-by-base view of the genome and therefore not only large variants but also small variants may be detected. Additionaly, there is promising news in the field of pharmacogenomics where information about the response or adverese effects of each individual to specific medications is predicted [3].

1.2 *De Novo* sequencing

De Novo sequencing approaches are used to sequence a new genome or transcriptome without any prior knowledge of the sequence where no reference sequence for alignment is given [5]. *De novo* methods are essential for mapping genomes when the genomes are not known but they are also extremely useful even when finishing genomes of known organisms [6]. Another desirable feature of *de novo* sequecing method is that the partial sequence can be used to search for posttranslation modifications or for the complex rearrangements, such as deletions, inversions, or translocations [7].

1.3 Metagenomic sequencing

Over the past few decades, metagenomics has become a standard tool to determine and study microbial communities of as yet non-cultivable microbes [8, 9, 10]. Metagenomic approaches enable comprehensive sequencing of all genes in all organisms present in a given complex sample obtained directly from an environment with no need to isolate and culture individual microbes [9]. In the classification and identification of bacteria, archaea or fungi present within a given sample, sequences of 16S ribosomal RNA gene are usually used. The 16S rRNA gene comprises nine variable regions interpresed between conserved regions, where conserved regions reflect phylogenetic relationship among species and highly variable regions determine differences between species [11].

1.4 Single-cell sequencing

In order to investigate structural and functional diversity and interactions in complex microbial ecosystems, as well as disease in multicellular organisms, the field of single-cell sequencing started to show its important potential as cells are the basic unit of an organism [12]. The thing is that every cell in our body contains nearly the same sets of genes, but transcriptome, on the other hand, reflects the cellular activity of only a subset of genes from the genome that are functionally active. In the case of bulk sequencing, many cells are sequenced together and consequently gene expression patterns at the population level are obtained [12, 13]. Therefore, the whole single-cell performance is evaluated from only single isolated cells to acquire expression at single-cell resolution.This strategy seems to hold great promise for sequencing of cells without prior knowledge of genes and proteins of interest as well as grouping of cells based on their transcriptional signature, which has been widely applied in the field of cancer biology, oncology, immunology or prenatal diagnosis [14, 15].

2 Generations of sequencing technologies

With the availability of sequencing technologies, study and analysis of nucleic acid composition for specific applications is accomplished and will be helpful in the area of basic science as well as translational research areas such as clinical diagnostics, metagenomics and forensic science. In recent years, the pace of progress has increased and novel techniques have been developed which leads to exponential reduction in cost per base. Furthermore, there are also many other important factors to consider such as read length, base per second and raw accuracy [16]. Tab. 2.1 outlines basic features and performances of the selected sequencing platforms. Accordingly, in this chapter, the three generations of sequencing technologies and the specifics on how a few different methods work, will be discussed.

Tab. 2.1: Basic features and performances of the selected sequencing platforms

	Run Time	**Output**	**Reads/Run**	**Read Length**
454 (Roche)				
GS FLX+	23 hrs	700 Mb	1M	up to 1 Kbp
GS Jr.	10 hrs	35 Mb	0.1M	700 bp
Illumina				
iSeq 100 System	9–17.5 hrs	1.2 Gb	4M	2 × 150 bp
MiniSeq System	4–24 hrs	7.5 Gb	25M	2 × 150 bp
MiSeq Series	4–55 hrs	15 Gb	25M	2 × 300 bp
NextSeq Series	12–30 hrs	120 Gb	400M	2 × 150 bp
HiSeq 4000 System	<1–3.5 days	1500 Gb	5M	2 × 150 bp
HiSeq X Series	<3 days	1800 Gb	6B	2 × 150 bp
NovaSeq 6000System	∼13–44 hrs	6000 Gb	20B	2 x 250
Ion Torrent				
PGM 314	2-4 hrs	200Mb	0.6M	400 bp
PGM 316	3-5 hrs	2Gb	3M	400 bp
GM 318	4-7 hrs	4Gb	5.5M	400 bp
PI	2-4 hrs	20Gb	82M	200 bp
PII	2-4 hrs	64Gb	330M	100 bp
SOLiD				
5500xl	6 days	95 Gb	800M	2 x 60 bp
5500xl Wildfire	10 days	240 Gb	2.4B	2 x 50 bp
5500	6 days	48 Gb	400M	2 x 60 bp
5500 Wildfire	10 days	120 Gb	1.2B	2 x 50 bp
PacBio				
RS II (P6-C4)	240 min	2 Gb	50k	10 -15 kbp
Sequel	240 min	20 Gb	500k	10 -15 kbp
Oxford Nanopore				
MinION	1 min–48 hrs	15-30 Gb	7 - 12M	entire fragment
GridION	1 min–48 hrs	15-30 Gb		entire fragment
PromethION	1 min–72 hrs	100-180 Gb		entire fragment
Flonge	1 min–16 hrs	1-2 Gb		entire fragment

2.1 The First Generation of Sequencing

The first techniques to be widely adopted and also considered as the real birth of first-generation DNA sequencing are Sanger's chain-termination method and Allan Maxam and Walter Gilbert's chemical cleavage or chain-degradation method, both developed in 1970s [17]. Both methods, also shown in Fig. 2.1, are described down below. The discovery of these techniques attracted interest of researchers and lead to development of faster and efficient sequencing technologies. Thus, a number of improvements upon existing methods were made which contributed to the development of increasingly automated DNA sequencing machines.

Sanger sequencing

Sanger sequencing, also known as chain-termination or dideoxy technique or sequencing by synthesis method was developed by Sanger et al. from Cambridge university awarded a Nobel Prize in chemistry in 1980 [18]. Until now Sanger sequencing has been considered as one of the most influential innovations that helped in a wide variety of biological researches. First of all, this well-established method requires a single-stranded DNA template. In order to use one strand of the double stranded DNA as template to be sequenced, the DNA needs to be denatured by heat so that the two strands separate. Denatured DNA template is then divided into four separate sequencing reactions, each of which contains primer, DNA polymerase, four deoxyribonucleoside triphosphates (dNTPs) and one of four chemically modified nucleotides called dideoxynucleoside triphosphates (ddNTPs) [18]. These radio- or fluorescently-labeled ddNTPs cannot form a bond with the 5' phosphate of the next dNTP due to a lack of 3' hydroxyl group. Therefore, once incorporated into the DNA strand they prevent further extension and the elongation is complete. Accordingly, as ddNTPs get randomly incorporated, strands of each possible length are produced and may be subsequently separated by the use of capillary electrophoresis [17]. While accuracy, robustness and ease of use are the main advantages of this method, it still sequences a single fragment at a time which makes this method not only time consuming but expensive as well [19].

Maxam-Gilbert sequencing

Maxam-Gilbert technique developed by Allan Maxam and Walter Gilbert, on the other hand, relies on the use of chemical reagents and thus is known as the chemical degradation method [20]. This chemical treatment modificates purified DNA and causes cleavage at a specific bases.

The DNA is radioactively labelled at one end and after the breakage of molecule at one or two predictable bases (G, A+G, C, C+T), series of marked fragments is generated [18]. These fragments are then size-separated using electrophoresis and can be subsequently visualized on exposed X-ray film [20]. By far the most important advantage is capability of directly sequencing purified double-stranded DNA and despite the usage of toxic and radioactive chemicals, the method has been widely applied for DNA footprinting [18].

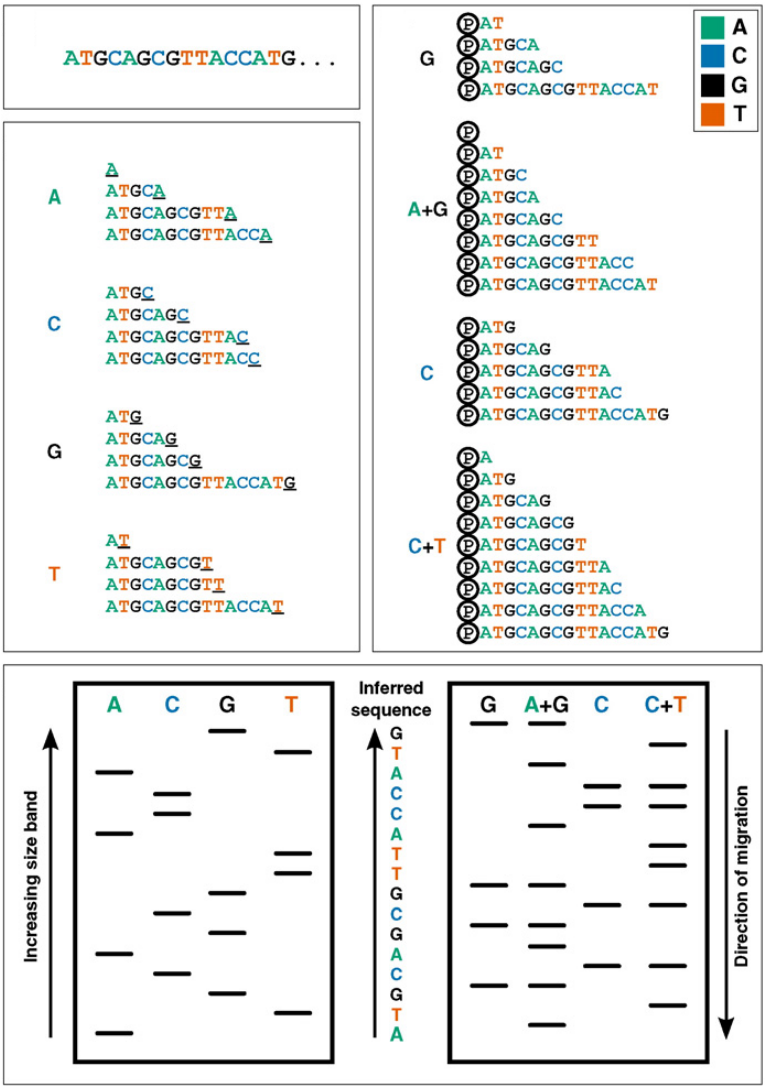

Fig. 2.1: Sanger sequencing (left) and Maxam-Gilbert sequencing (right) [17]

2.2 Next-Generation Sequencing

Next-Generation Sequencing (NGS) is a powerful tool that has enabled parallel sequencing of thousands to millions of DNA molecules simultaneously and is considered as one of the most influential technological advances in the biological sciences of the last few decades. Compared to other generations, this generation is attracting considerable interest due to its sensitivity, speed, and reduced cost per sample [21]. Recently, NGS has been used by an increasing number of researchers for de novo genome sequencing, DNA resequencing, transcriptome sequencing and epigenomics [4, 6, 9, 12]. To clearly understand the evolution of sequencing technology from the first generation sequencing, the second and the third generation of sequencing will be discussed separately in more detail. An overview diagram shown in Fig. 2.2 presents a hierarchical structure of the corresponding methods .

Fig. 2.2: Overview of the Next Generation Sequencing

2.2.1 The Second Generation of Sequencing

Second-generation sequencing have been made available on an increasing range of platforms designed to suit different applications and capacity requirements as well. With these technologies, to achieve massive parallel sequencing, it is necessary to clonally amplify DNA templates on a solid surface or on beads while isolated within miniature emulsion droplets or arrays [22]. Moreover, the advance of second-generation technology has been enabled by innovations in monitoring nucleotide incorporation, such as luminescence detection or detection by changes in electrical charge during sequencing procedure [23]. Some of the major and commonly utilized sequencing platforms will be briefly described in this section.

Roche/454 sequencing

Pyrosequencing is method that utilizes two-enzyme process required for the sequencing-by-synthesis approach, the same principle as Sanger's dideoxy method relies on. However, this technique detects the activity of DNA instead of the detection of radio- or fluorescently-labelled nucleotides [24].

Pyrophosphate is detected by enzyme cascade reaction, as shown in Fig. 2.3, that results in the emission of light [25]. The emission of light confirms that a pyrophosphate has been released. When this pyrophosphate combines with another substrate known as Adenosine Phosphosulphate (APS) in the presence of an enzyme ATP sulfurylase, ATP is generated. In the next reaction this ATP is utilized by the enzyme luciferase for the conversion of lucifer into oxyluciferin and production of light. Thus, the pyrophosphate released during DNA synthesis can be detected by the emission of light. [24]

To begin with sequencing, DNA samples are randomly fragmented and then attached to beads via adapter sequences [18]. This DNA fragment serves as DNA template strand and it is incubated with the primer binds to its complementary sequence on the DNA template strand. In the next step, DNA polymerase is added along with the enzymes and substrates required for the detection of the pyrophospate [23]. After that, one of the four types of nucleotides is added and only one type of nucleotide is added at a time. If the added nucleotide is incorporated in the new strand, pyrophosphate will be released and emission of light take place. This light is detected by a detector and later used to interpret unknown sequence. After the degradation of unused and extra nucleotides by added enyzme apyrase, the reaction starts again with the addition of next nucleotide. This process is repeated adding each nucleotide one after another until the sythesis si complete. The amount of light generated is proportional to the number of nucleotides that are incorporated [26].

The light emission is then represented graphically to interpret the sequence [25]. The peaks in the graph also give an idea about the number of same nucleotides present in the sequence.

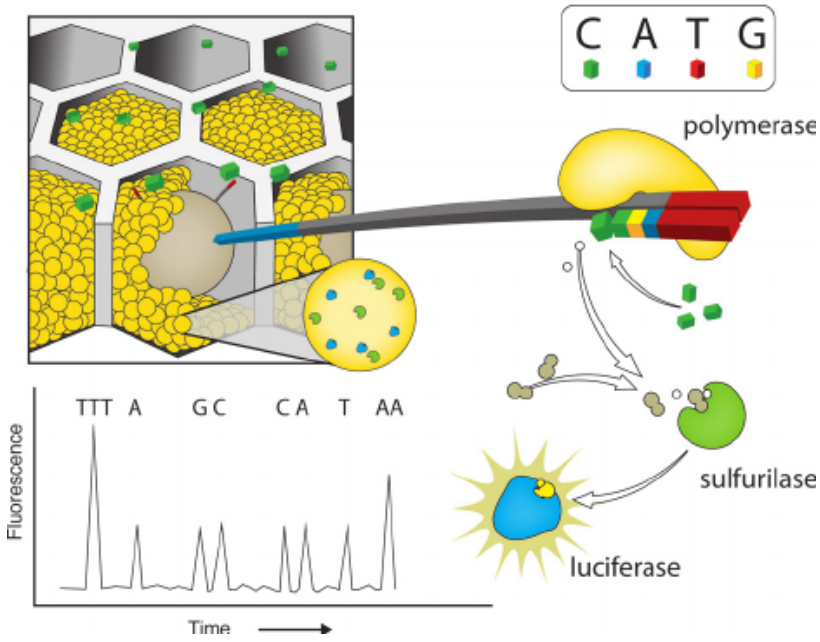

Fig. 2.3: Pyrosequencing method [25]

Ion Torrent sequencing

Unlike other sequencing technologies, even with its similarities to pyrosequencing technology, Ion Torrent Systems sequence DNA using a semiconductor chip, which has millions of wells [18]. These wells capture chemical information from DNA sequencing and translate it into digital information or base calls. The sequencing process starts when a sample of DNA is cut into millions of fragments. Each fragment then attaches to its own bead and is copied until it covers the bead. This automated process covers millions of beads with millions of different fragments [25]. These beads then flow across the chip, each depositing into a well. Then the chip is flooded with one of the four DNA nucleotides. Whenever a nucleotide is incorporated into a single strand of DNA, a hydrogen ion is released. The hydrogen ion changes the pH of the solution in the well. A sensor attached to the bottom of the well measures that change in pH and converts it to voltage [22]. This voltage change is recorded indicating that the nucleotide was incorporated and the base was called. In essence, each well works as the world's smallest pH meter. The process happened simultaneously in millions of wells and is repeated every 15 seconds with a different nucleotide washing over the chip [24].

Illumina/Solexa sequencing

With Illumina sequencing platform, the first step after DNA purification is random fragmentation and ligation of adapters to both ends of each sequence followed by reduced cycle amplification [25]. Through this cycle, additional motifs are indtroduced, such as sequencing binding sites, indices and regions that are complementary to the flow cell oligonucleotides. DNA fragment strands with adapters are subsequently loaded into a flow cell channels, where two types of mentioned complementary surface-bound oligos are placed. Once attached, every single strand is then amplified by PCR bridge amplification, as Fig. 2.4 shows, in which strand folds over and the adapter region hybridizes to the second type of oligo on the flow cell [23]. A DNA polymerase synthesizes the complementary strand resulting in double stranded bridge. The double stranded DNA is denatured and reverse strands are cleaved and washed off leaving only the forward strands [22]. Several milion dense clusters of sequences made from the same original sequence are generated in each channel of the flow cell. Each cluster act as an individual sequencing reaction where reversible terminators in which the four modified nucleotides, sequencing primers and DNA polymerases are added as a mix to the flowcell [18]. This process, also known as sequencing by synthesis, begins with the extenstion of the attached primer to the DNA being sequenced. The fluorescently tagged nucleotides compete for addition to the growing chain and also have an inactive 3'-hydroxyl group, so that only one base is incorporated at a time. Once a base has been added the clusters are excited by a light source and a characteristic fluorescent signal is emitted.

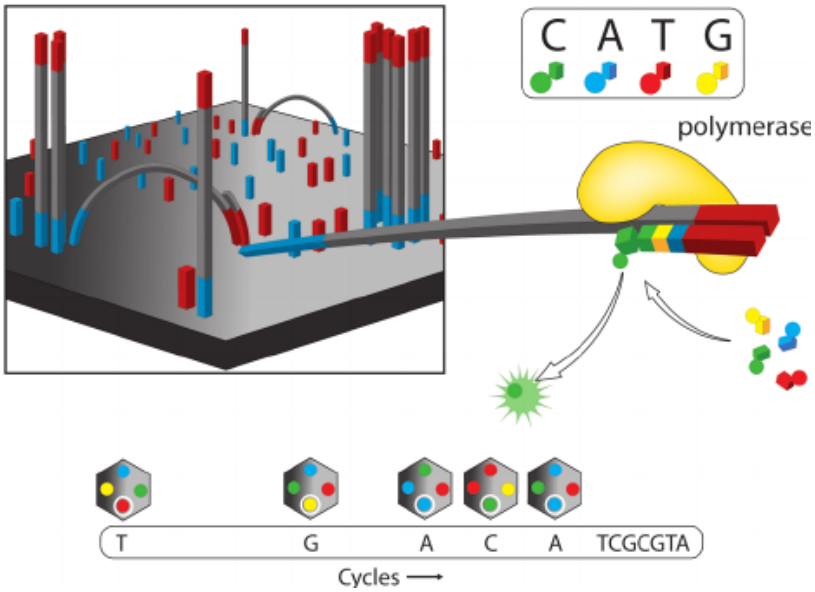

Fig. 2.4: Illumina sequencing method [25]

Afterwards, this fluorescence is detected by a CCD camera and using computer programs these signals are converted into a nucleotide sequence [26]. The process continues with the elimination of the terminator with the fluorescent label and the starting of a new cycle with a new incorporation. In order to determine each nucleotide in the sequences, the terminator with the fluorescent label is removed and whole process is repeated with the next fluorescently labelled base until millions of clusters have been sequenced [24].

ABI/SOLiD sequencing

The process starts by attaching adapters to the DNA fragments. The fragmentation can be achieved in one of three ways - nebulization, sonication and digestion, causing the DNA to shear at random intervals [27]. Clonally amplified DNA fragments to be sequenced are then linked to magnetic beads [25], as shown in Fig. 2.5. These beads and fragments are then put in an emulsion, so that small units of beads and fragments are formed [18]. The beads are then chemically bound to a glass plate. Each plate contains millions of beads, each with a specific DNA fragment at a specific position. The next step is to sequence all the beads in parallel. This method uses DNA ligase to generate DNA sequence by measuring the serial ligation of an oligonucleotide to the DNA. The sequencing starts by attaching a primer to the adaptor. Next, a probe that interrogates two bases and has a fluorescent dye linked to it, is hybridized, ligated to the template using mentioned DNA ligase, and detected by fluorescence imaging [25].

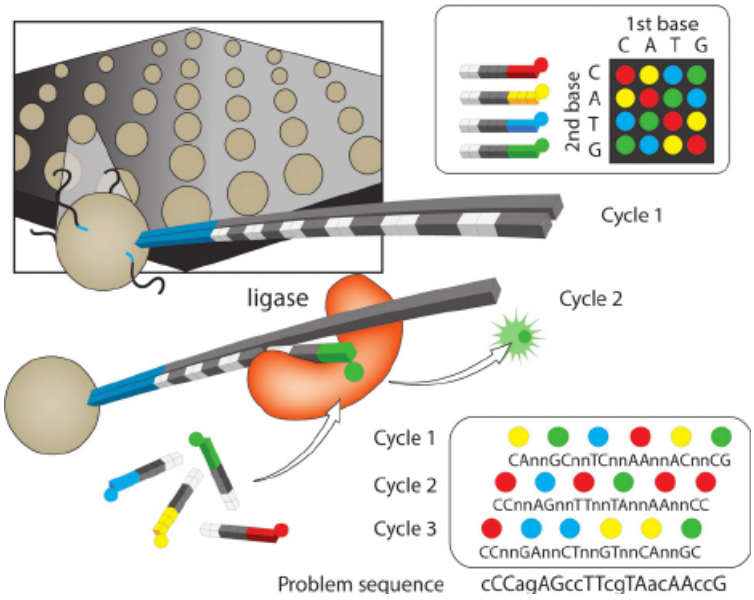

Fig. 2.5: SOLiD sequencing method [25]

There are 16 different two base combinations and each base is interrogated twice by two different dye-labeled probes. Finally from the color that the beads emit, it is obvious what group the first two bases of the DNA strand on each bead belong [25]. Corresponding color is stored for each bead and after reading the first color, the fluorescent dye label can be cleaved thereby preparing the system for another round of ligation [27]. In order to get color codes for the other positions, the whole sequencing process is repeated to sequence the complete target DNA. However, each time the next round of sequencing is performed with a primer that is one base shorter to sequence skipped positions. At the end, to obtain required DNA sequence, the recovered data from the color space are translated to letters [24].

2.2.2 The Third Generation of Sequencing

Third-generation sequencing technologies offers many theoretical benefits such as reduced cost and preparation time, increased speed and eliminated PCR-biases and errors [23]. The main difference with second-generation sequencing is the shift to single-molecule PCR-free protocols and cycle-free chemistry so that no clonal amplification is required, which make this technology has the potential of becoming one of the most promising platforms [22]. However, the technologies are still at very different stages of development, some of which have already launched and some of which are still in stealth mode so it may take a long time to become fully functional and widely available.

Pacific biosciences SMRT sequencing

Pacific Biosciences Single-Molecule, Real-Time (SMRT) sequencing technology, most widely used third-generation method, enables the observation of DNA synthesis as it occurs in real time. Sequence information is captured during replication process of the template to be sequenced. The single-stranded closed circular DNA, also called a SMRTbell, is created by ligating hairpin adaptors to both ends of a target double-stranded DNA [23]. The method uses fluorescent labelling, but in contrast to other sequencing approaches, phospho-linked nucleotides carry their fluorescent label on the terminal phosphate rather than the base [16]. Fluorescent label, as part of the incorporation process, is then cleaved away resulting in completely natural strand of DNA. A single polymerase is immobilized at the bottom of the chamber called Zero Mode Waveguide (ZMW), shown in Fig. 2.6, where the target DNA fragment is placed [28]. The ZMW is a cylindrical metallic chamber approximately 70 nanometers wide and it enables observation of individual molecules against the required background of labelled nucleotides.

Whenever one of four fluorescent-labelled nucleotides is incorporated, distinct emission spectrum is generated and subsequently captured by a sensitive detector. Nucleotides diffuse in and out of the ZMW and after incorporation, the label is clipped off and diffuses away. In order to determine the DNA sequence, the whole process repeats creating sequential bursts of light corresponding to the different nucleotides [29]. The main advantage of this method is that it offers much longer read lengths and faster runs than SGS methods. Additionally, it allows simultaneous multiplexing of thousands of ZMWs in parallel, all concurrently replicating DNA in real time. However, lower throughput, higher error rate and higher cost per base appear as main disadvantages of this method [18, 30].

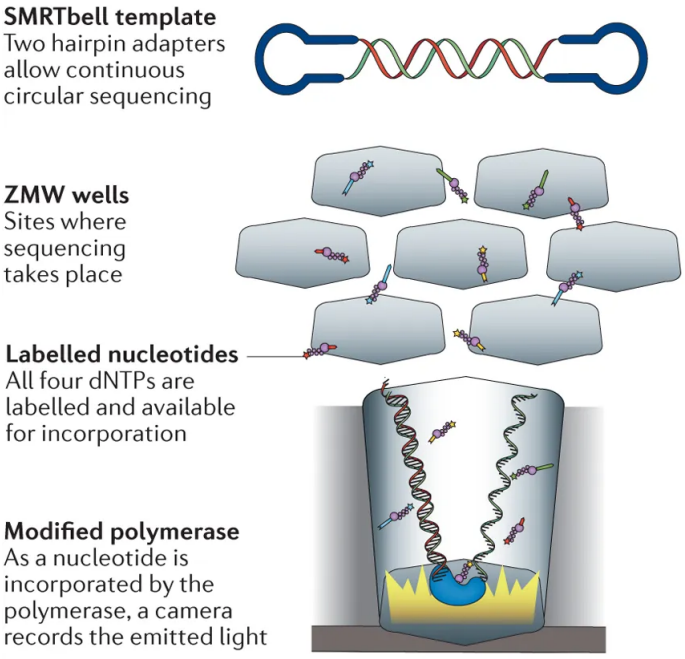

Fig. 2.6: Pacific biosciences sequencing [31]

Oxford Nanopore Technology sequencing

Nanopore sequencing technology offers direct, real-time analysis of DNA or RNA molecules. This technology shows greater promise compared with other sequencing techniques, because of its portability, long reads and ease of set up by those with fewer lab skills [30]. In order to sequence both strands of long double-stranded DNA (dsDNA), the library usually contains two adapters, the leading adapter and the hairpin adapter, both ligated to one end of the dsDNA [32].

The DNA strands to be sequenced are then mixed with copies of a processive enzyme, which is loaded at the 5'-end of the leading adapter. As the DNA enzyme-complex approaches the nanopore, the enzyme unzips the dsDNA and the single-stranded DNA is pulled through the aperture of the nanopore. After the template strand is sequenced, hairpin adapter is reached and followed by the complementary strand, for which sequencing process repeats. The nanopore inserted into an electrically resistant membrane plays a key role in this method [33]. A voltage can be applied across the membrane to drive DNA through the pore. These single molecules that enter the nanopore cause a characteristic disruption in the electrical current [30]. The current is measured by a sensor several thousand times per second and the corresponding information is then used to determine the order of the bases on that DNA strand. The so-called "squiggle plot", shown in Fig. 2.7, shows the raw current measurements over time, which can be subsequently translated into DNA bases [32].

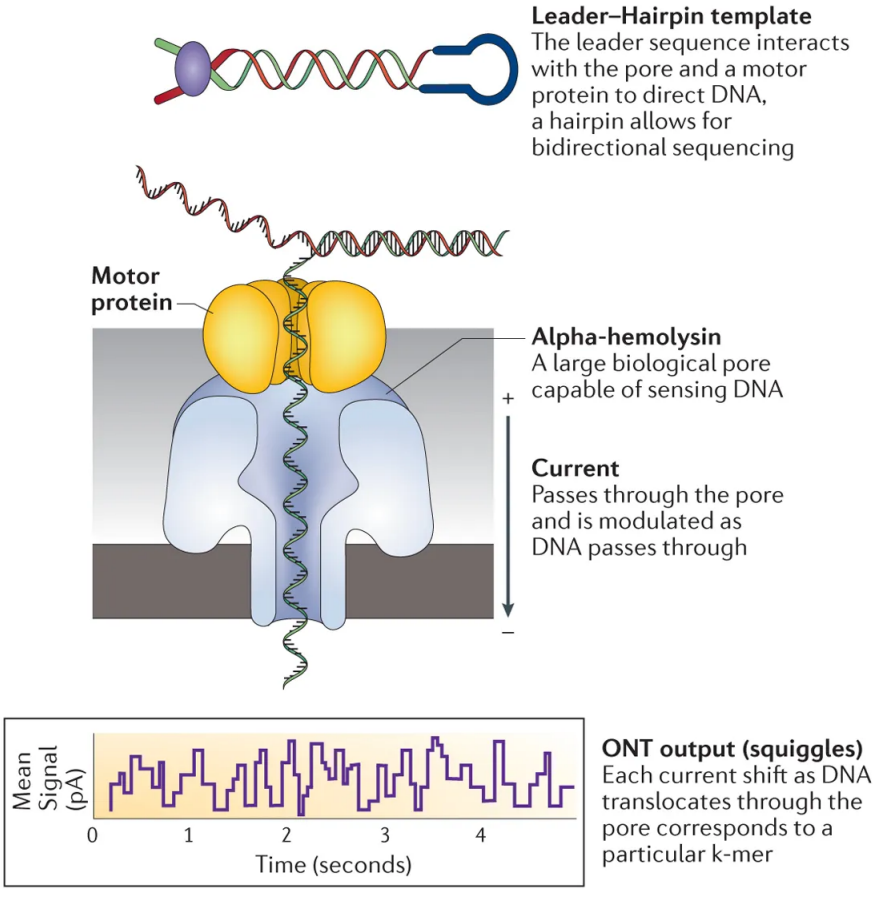

Fig. 2.7: Oxford Nanopore sequencing [31]

Heliscope single molecule sequencing

This sequencing technology utilizes sequencing-by-synthesis methodology, involving a DNA samples that are cut into short strands to which ends polyA tails are then attached [26]. These DNA strands are hybridized to the Helicos flow cell surface coated with oligo-T universal capture sites [34]. After each individual template hybridizes to the flow cell, generating its own sequencing reaction, a laser illuminates the surface of the flow cell showing the location of each fluorescently labelled template. After the incorporation event, a CCD camera images the the entire surface to produce a map of the templates on the flow cell surface. Once the templates have been imaged, the label is cleaved and washed away allowing to start the whole sequencing process by adding DNA polymerase and another fluorescently labelled nucleotides to be incorporated [34]. After the fluorescent label incorporation event, images are captured again and excess DNA polymerase and nucleotides are washed away. The process continues through each of four nucleotides, where images from each incorporation are analyzed, and repeats until the desired read length is achieved [16]. The main advantage is the lack of amplification steps and capability of generating accurate reads on captured fragments. On the other hand, each run need 14 terabytes of computer storage, however those 14 terabytes can hold an enormous amount of sequenced data [35].

GnuBIO sequencing

This droplet-based DNA sequencing platform combines microfluidic and emulsion technology, which effectively reduce the number of library preparation steps [24]. Thus, the whole process, including target selection, DNA amplification, DNA sequencing and analysis, is integrated into a single high throughput system [30]. Moreover, each mini-droplet works as a unique sequencing reaction including PCR amplicons and one of approximately 5000 labelled hexameric sequencing primers along with DNA polymerase [24]. Hybridization of a particular hexamer to a given amplicon is then observed and corresponding fluorescence is detected to determine which hexamers do or do not hybridize and, hence, this so called displacement reaction, serve to mark the presence or absence of the signal from the sequenced molecule that it would be possible to map the final sequence and its structural irregularities [30]. Using this scalable sequencing reaction, genomic results can be produced within hours.

3 RNA-sequencing

Over the past few years, RNA sequencing (RNA-seq) [36] become very powerful sequencing technique for transcriptome-wide analysis that utilizes next-generation sequencing platforms. It allows to reveal the presence and quantity of RNA in a biological sample at a given moment. Usually, mutated cells are analysed in order to discover what genetic mechanism is causing its different behaviour when comparing to normal cells. At this point, it is crucial to look at differences in gene expression. By analyzing the continuously changing cellular transcriptome, a better understanding of how gene expression can determine cell fate is accomplished. The recent development of novel and effective NGS methods has provided an ideal environment to develop new methods for both mapping and quantifying transcriptomes. High throughput sequencing tells us which genes are active, and how much they are transcribed. Typically, RNA-seq is used to measure gene expression in normal cells and mutated cells. These cells are then compared in order to figure out what is different in the mutated cells. In general, there are three main steps for RNA-seq: preparing a sequencing library, sequencing itself, and final data analysis.

3.1 Library preparation

As shown in Fig. 3.1, to prepare library for Illumina short-read RNA-seq (black line) [37], isolated RNA sequences are firstly sheared into small fragments. Secondly, RNA sequences are converted into the fragments of double stranded DNA. The reason is that DNA sequences are more stable and can be easily amplified and modified. In the next step, sequencing adaptors are added to generated DNA sequences to make it possible for the sequencing machine to recognize the fragments. Another advantage of these adapters is that different samples can be sequenced at the same time, since different samples can use different adaptors.

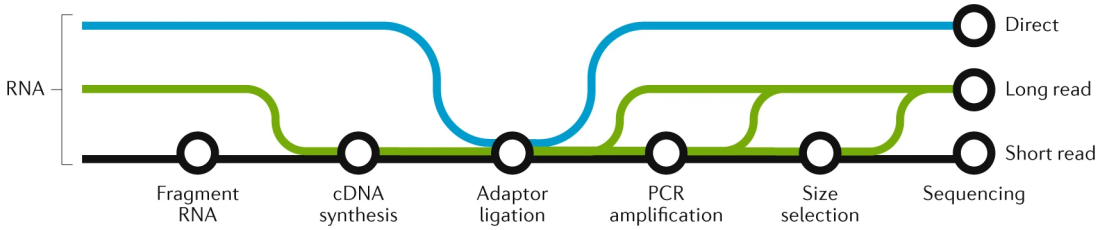

Fig. 3.1: Library preparation methods for different RNA-sequencing methods [36]

In the subsequent stage, PCR amplification of the library is carried out and only the fragments with sequencing adapters are amplified. At the end of the library preparation, the library concentration and the library fragment lengths are verified.

However, there is also alternative long-read (green line) or long-read direct (blue line) method that provides some advantages, such as reduction of ambiguity in the mapping of read sequences, identification of longer transcripts or reduction in detection of false-positive splice-junction [36, 38]. In case of Iso-Seq protocol [39], high-quality RNA is converted to full-length cDNA, which is then PCR amplified and used as the input for PacBio single-molecule, real time (SMRT) library preparation. In order to reduce bias in the sequencing of short transcripts, size selection of transcripts from 1 to 4 kb is usually performed. This short transcripts, which typically tend to diffuse more easily to the active surface of the sequencing chip, are then more equally sampled with considerably longer transcripts. Full-length transcripts are also generated by Oxford Nanopore technology (ONT) cDNA sequencing [38]. To prepare sequencing library, full-length cDNAs are optionally amplified before adaptor ligation. When no amplification is involved in the library preparation, PCR bias is avoided. On the other hand, PCR amplification is still very useful as it enables users to start with a much smaller amount of starting material. This is a trade-off that needs to be considered in the library preparation for each case of RNA-seq analysis. There is also mentioned long-read direct method (blue line), another nanopore sequencing demonstrated by ONT [40], without need to convert RNA to cDNA before sequencing. Library preparation does not require any cDNA synthesis nor PCR amplification and RNA, therefore, can be sequenced directly after adapters ligation. The whole library preparation process for this method is described in the previous chapter.

3.2 Sequencing process

After the library is prepared, it is then sequenced. The produced raw read sequences are usually in a FASTQ file format because, at first, low quality reads need to be filtered out. In general, reads with low quality base calls or obvious artifacts of the chemistry, when only the adapters bind to each other, are considered as low quality reads. The remaining high quality reads are then aligned to a genome. At first, a genome sequence is split into small fragments and afterwards, index of all the fragments and their locations within the genome is created. When analysed sequenced read is obtained, it is then also necessary to split the read into fragments. The splitting step is required in order to match the read fragments to the genome fragments. The aligning step is carried out by one of the available tools that perform a spliced alignment allowing for gaps in the reads when compared to the reference genome.

The reason is that cDNA derived from RNA may contain exon-exon junction which cannot be contiguously mapped to the genome. As shown in Fig. 3.2, it is essential to perform Quality Control (QC) [41] and look at the percentage of reads mapped to the reference genome, as it can indicate some issues with the data. Once the read fragment match the genome fragment, it is then easy to determine its location in the genome. By this, with known chromosome and position for a read, number of reads can be counted per each individual gene. However, different number of reads can be assigned to each sample, and therefore, data are usually normalized before downstream analysis.

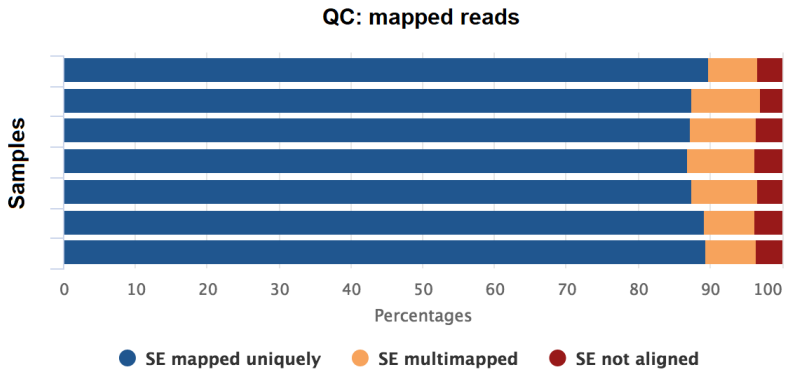

Fig. 3.2: Quality control of mapped reads [41]

As mentioned previously, RNA-seq data usually undergo PCR amplification step resulting in high duplication rates. In order to improve the quantification of gene expression and the estimation of allele frequency, random Unique Molecular Identifiers (UMIs) are added to cDNA molecules before amplification [42]. In the case of RNA-seq, duplicate reads are considered as an indication of a true biological signal and UMIs have been proposed as the best way to distinguish technical from biological duplication. As shown in Fig. 3.3, after the alignment, UMI deduplication takes place before previously described quantification.

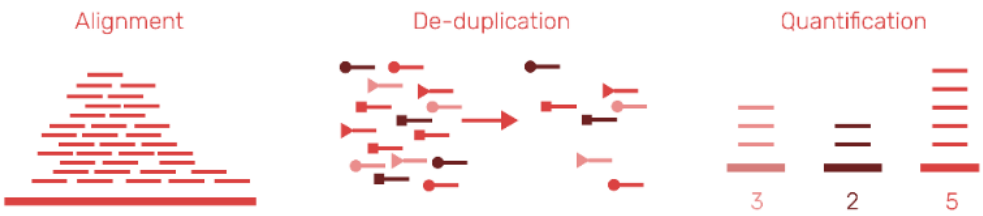

Fig. 3.3: Elimination of PCR duplicates in RNA-seq

4 Unique Molecular Identifiers

Not only RNA-seq methods, but many other high-throughput sequencing platforms require PCR amplification to be performed before sequencing [22]. Nevertheless, different molecules are amplified with unequal probabilities so this step can easily lead to certain sequences becoming excessively presented in the final library. Therefore, random oligonucleotide barcodes, otherwise referred to as UMIs, have been used to distinguish between identical copies arising from distinct molecules and PCR duplicates arising from the same molecule [43]. Duplicate sequencing reads produced by PCR amplification may lead to mentioned biases which reduce quantitative accuracy and cause misleading interpretation of sequencing results [44]. However, the problem of PCR duplicates is more acute as sequencing depth increases and reads or read pairs with the same alignment coordinates are removed even if they originated from two different molecules, or when greater numbers of PCR cycles are required to ensure sufficient DNA for sequencing so to increase the library concentration [45]. Moreover, a distinct identity for each input molecule established by attached UMI makes it possible to identify sampling bias and estimate the efficiency with which input molecules are sampled [46]. Many tools have been used to perform deduplication of sequenced reads by their UMIs, and therefore, in the following sections, an overview of some of the most used tools will be presented.

4.1 UMI-tools

UMI-tools [43] contains tools for dealing with UMIs and single-cell RNA-Seq cell barcodes. For accurate quantification with UMIs, the number of unique UMI barcodes at a given genomic locus and the number of unique fragments that have been sequenced, are taken into consideration. The problem is that during PCR or during sequencing, UMI errors, such as nucleotide substitutions, deletions or insertions, have been detected. In fact, these errors within the UMI sequence create additional artificial UMI and therefore the number of unique molecules at a particular genomic coordinate can be overestimated, thus quantification accuracy is negatively affected. As shown in Fig. 4.1, different methods were employed to resolve UMI errors by examining all UMIs at a single locus. One well-known method to identify unique molecules is called unique and assume that each UMI at a given genomic locus represents a different unique molecule. Otherwise, the percentile method considering sequencing error issues attempts to remove UMIs at a given locus whose counts fall below a threshold of the mean of all nonzero UMIs. In addition, three other methods have been developed by this tool.

All of the methods work with UMI networks formed by linking UMIs separated by a single edit distance. These networks are afterwards reduced to obtain representative UMIs. The first one, called cluster, merges all UMIs within the network to keep only the UMI with the highest count. The number of unique molecules is then the same as the number of networks formed at a given genomic position, however the method usually underestimates the number of unique molecules for complex networks. Because of that, the adjacency method in which it is possible that a complex network originates from more than one UMI, has been developed. It works with the node counts to iteratively remove the node with the highest abundance and its neighbours from the network. Even though UMIs with an edit distance of two between any two nodes are removed individually in two different steps. The number of predicted unique molecules with the same genomic coordinates is equal to the number of steps to resolve the network formed at this specific locus. In the third and final method, called directional, networks consist of nodes and directional edges that connect nodes a single edit distance apart if the counts of the first node is approximately two times greater than the counts of the second node. The node with the highest count is then considered the key node from which the network originated. In order to estimate the number of unique molecules, the number of directional networks formed is observed.

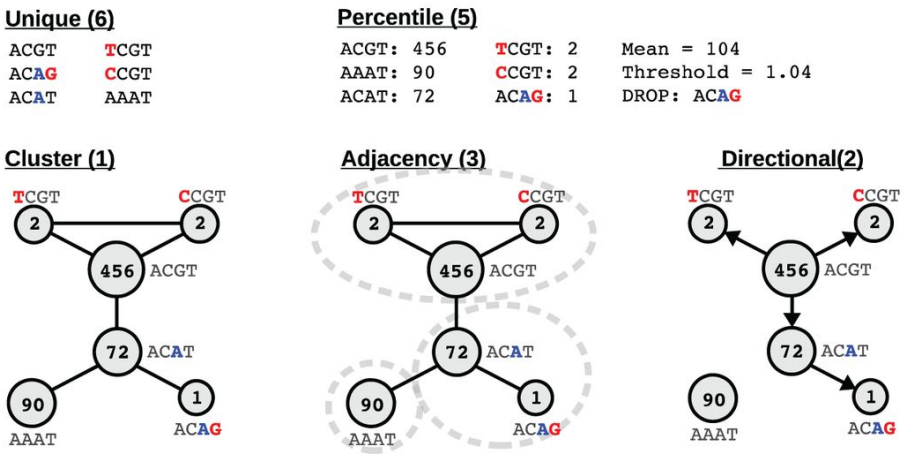

Fig. 4.1: Methods for estimating unique molecules (red bases - sequencing errors, blue bases - PCR errors, () - number of estimated unique molecules) [43]

4.2 Gencore

Gencore [47] is a tool, which is useful for performing deduplication and consensus read generation for NGS sequencing data. According to authors, a comparison with Picard [48] , Samtools [56] and UMI-tools [43] showed that the tool is much faster and more memory efficient with similar or even better results. This tool, as well as UMI-tools, does not require any additional BAM preprocessing before performing deduplication [43]. Besides this position sorted BAM file, a FASTA file as reference genome is also required as an input. As illustrated in the workflow shown in Fig. 4.2, all mapped read pairs are firstly grouped by mapping position where in each group, read pairs are then clustered by their UMIs. Formed clusters are then filtered by its supporting reads number. For the remaining clusters, a consensus reads are generated and overlapped region of the paired reads for each read pair in a cluster is computed. With the consideration of the quality scores, each base in the overlapped region is scored according to its paired base. These scores are then summarized to obtain a total cluster score. On top of that, base diversity for each position in the mapping region is computed and most frequently represented bases are then assembled to generate consensus read. As a result, an output in HTML or JSON format reporting particular metrics is generated.

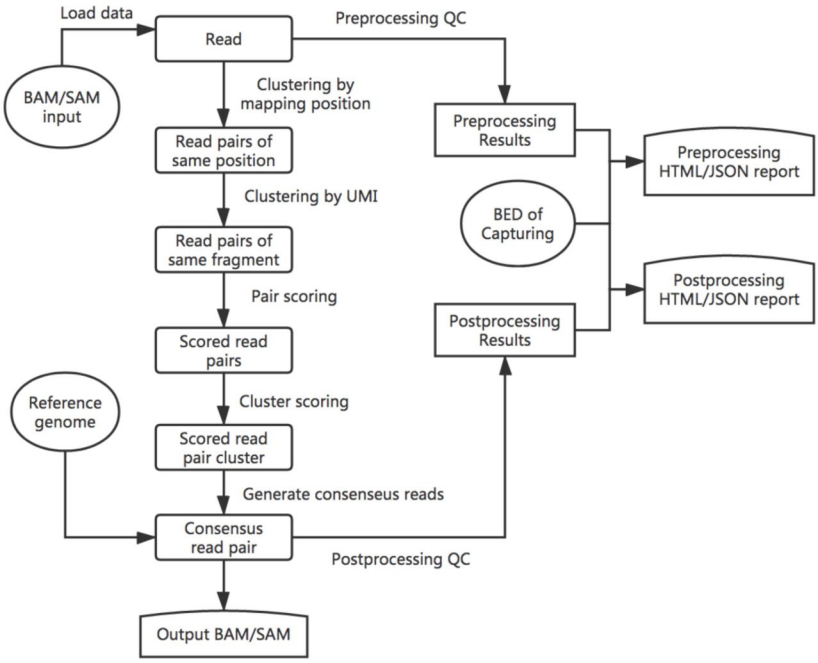

Fig. 4.2: Schematic of the gencore pipeline [47]

4.3 UMI-Reducer

UMI-reducer [49] was also developed to differentiate technical duplicates, which are collapsed to a singe unique read, from biological duplicates, which are considered as separate unique reads. However, tool is only suitable for RNA data [47], the final set of all unique reads seems to be very effective in the estimation of mRNA abundance. As shown in Fig. 4.3, to obtain BAM-formatted alignment file that includes biologically unique reads with their mapping positions, raw reads are firstly mapped to reference genome. Secondly, the BAM file is then analyzed to identify reads that are mapped to the same position in the genome. Additionally, if these reads have identical UMIs, they are categorized as PCR duplicates and afterwards are collapsed into a single read. The final set of reads annotated to the genomic region can be easily used to count the number of reads per gene in each genomic region.

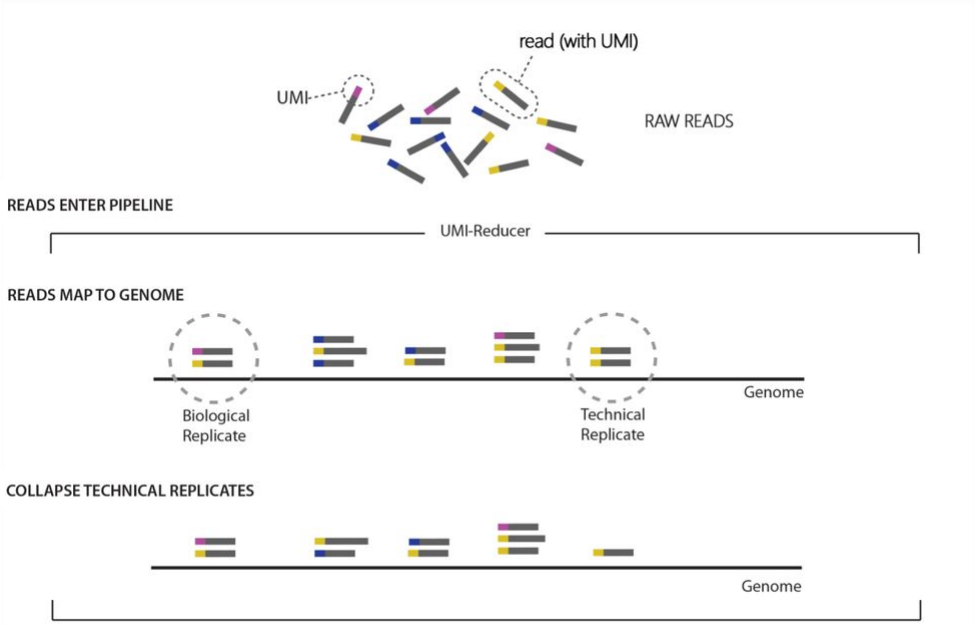

Fig. 4.3: UMI-Reducer pipeline [49]

4.4 zUMIs

In the zUMIs pipeline [50], shown in Fig. 4.4, the first step is filtering, where according to a user-defined threshold, reads that have lower quality UMIs/BCs are removed. Using the splice-aware aligner STAR [51], the remaining reads are then mapped to the genome. Using Rsubread featureCounts [52], reads are assigned to genes based on two annotation files from gtf with provided exon and intron positions. The output is then read into R, generating count tables for UMIs and reads per gene per BC. In addition, several data and plots for quality measures are generated.

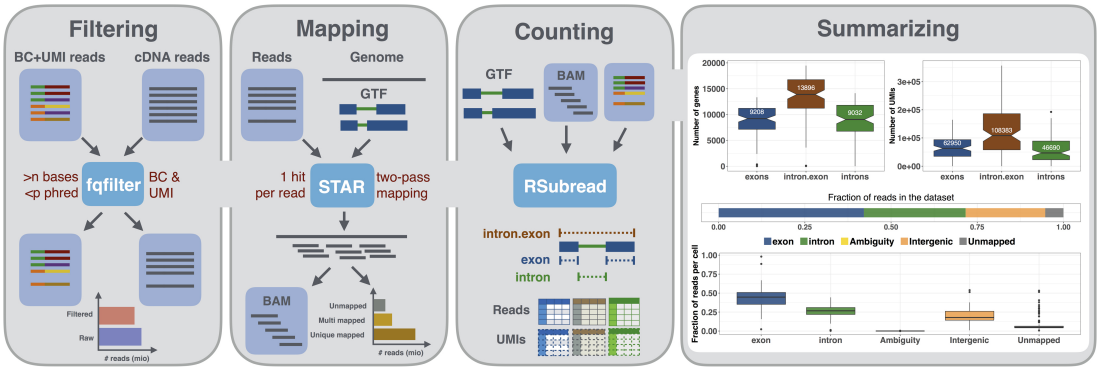

Fig. 4.4: Schematic of the zUMIs pipeline [50]

4.5 Other methods

Picard MarkDuplicates

Picard MarkDuplicates [48] takes as an input BAM or SAM file and locates and tags duplicate reads by comparing sequences in the prime positions of both reads and read pairs. By the sums of reads base-quality scores, collected duplicate reads are distinguished into primary and duplicate reads. As a result, new SAM or BAM file with identified duplicates together with metrics file containing numbers of duplicates for both single- and paired-end are obtained as the output.

fgbio

Fgbio [53] also provides a set of well-tested tools to analyze genomic data, especially tools for manipulating UMIs or reads tagged with UMIs. This tool makes it also possible to group reads together that originated from the same molecule.

Similarly to UMI-tools [49], reads with the same mapping positions are grouped by one of three methods. In the first one, called identity, only reads with the same UMIs are grouped together. The second one, edit method, clusters reads that are within specific edit difference from each other. The last but not least is the adjacency method, which is pretty much the same as described in UMI-tools [43] section. Then there is the paired method, which is very similar to previous one, however more preferred when template with pair of UMIs is produced.

Je

Je [54] tool has the ability to extract UMIs from reads and filter duplicates with or without pre-defined list of UMIs. Besides that, it also offers the remarkable ability to handle mismatches in the UMI sequence during filtering. Moreover, complex barcoding configurations are supported, such as barcodes inserted at each fragment end in paired-end sequencing used to sample multiplexing or to use one of the barcodes as UMI. These barcodes are useful when large numbers of libraries are pooled and need to be sequenced in a single run to make next-generation sequencing as efficient and affordable as possible [55]. Basically, Picard's MarkDuplicates tool [48] is used to identify PCR read duplicates based on their mapping positions, UMIs, and chosen scoring strategy.

5 Implementation

This section describes an automated bench-marking pipeline created to compare performance of our algorithm called My_UMI_tool with the existing one called UMI-tools [43]. The proposed pipeline is presented graphically in Fig. 5.1 as a flowchart. The UMI-tools was selected for comparison as it gives a good representation of the methodology used in quantitative research and has relatively high prediction accuracy. Moreover, because of its effectiveness, it is one of the most commonly used tools. Thus to obtain the total number of unique molecules becomes a challenging task, especially when the purpose is to obtain efficiency and effectiveness improvements. At first step in the proposed pipeline, the user selects a FASTQ-format sequence files of interest for further processing. These files contain multiple sequences including UMI barcodes and developed computational pipeline simultaneously processes them by two different approaches as illustrated by the blue and red arrows in Fig. 5.1. Moreover, the user can select one or more files at once, which indicates that the number of final output files changes depending on the number of input files.

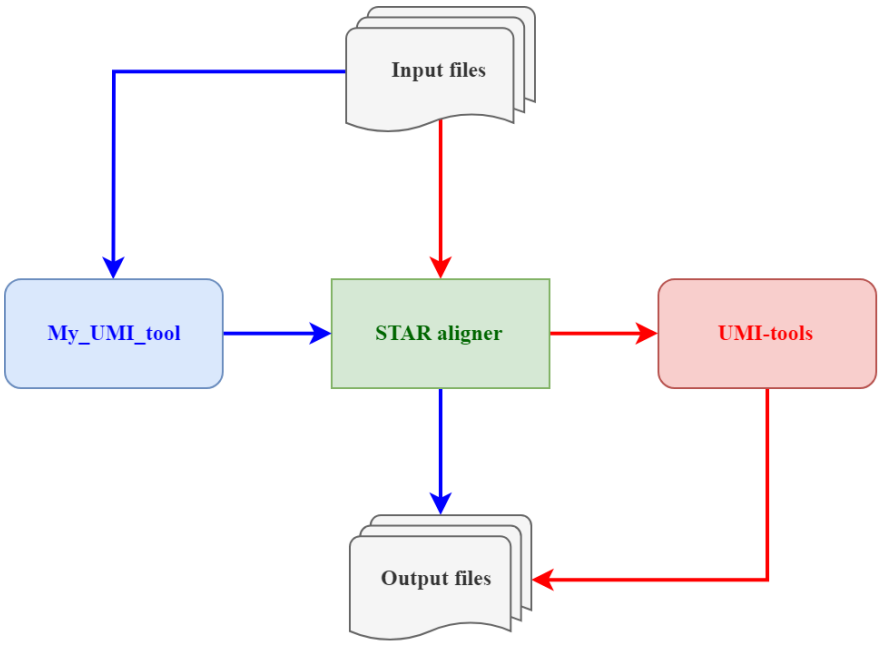

Fig. 5.1: Schematic diagram of the proposed pipeline

After the selection of input files, there are two ways to proceed from here. Reads from input files could be mapped to the genome and assigned to genes first, using STAR aligner [51], and then proceeded by UMI-tools. Otherwise, input FASTQ files could be proceed by our algorithm My_UMI_tool first and remaining reads mapped

to the reference genome. Before the proposed approaches will be described in detail, Tab. 5.1 includes a short overview to help with orientation and understanding of the analysis phases and included input and output files for each individual tool involved in the proposed pipeline.

Tab. 5.1: Input and output files for involved tools in proposed pipeline

	Input files	**Output files**
My_UMI_tool	FASTQ (Reads)	FASTQ TSV
STAR alignment	FASTA (Reference sequence) GTF (Annotations) Index file FASTQ (Reads)	BAM BAM.BAI
UMI-tools	BAM	BAM BAM.BAI TSV

5.1 STAR aligner

For accurate alignment, fast universal RNA-seq aligner, called STAR [51], is used. This aligner was designed to align the non-contiguous sequences directly to the reference genome. STAR outputs aligned sequences in BAM files, compressed binary version of a SAM file, sorted by coordinates. There is also a BAM index file (BAI), which provides an index of the corresponding BAM file.

Generally, the aligner involves two basic steps. Firstly, genome index files are generated, and secondly, reads are mapped to the genome. The output of the first step is genome index file, generated from input FASTA file of the reference genome sequence and annotation file. In particular, these two files have to match chromosome names. However, these indexes need to be generated just once for each combination of genome and annotation. Therefore, in our case, index file is loaded from the disk. In the second step, the output index file of the first step is combined with input reads (sequences) in the form of FASTQ file to finally map reads to the genome and write output BAM file. The mapping algorithm itself includes another two steps. The first one, seed searching step is the sequential search for a Maximal Mappable Prefix (MMP). To obtain MMP, read sequence, read location, and a reference genome sequence must be given.

The MMP refers to the longest substring of read sequence that corresponds to one or more exact substrings of a reference genome sequence. As shown in Fig. 5.2 if there is an exon-exon junction in a read sequence, it cannot be mapped contiguously to the genome. At this point, the seed searching step is applied, and thus, the first seed is mapped to a donor splice site. The unmapped portion of the read is subsequently processed by the MMP search again and afterwards is mapped to an acceptor splice site. By this approach, not only splice junctions are identified, but also mismatches and indels.

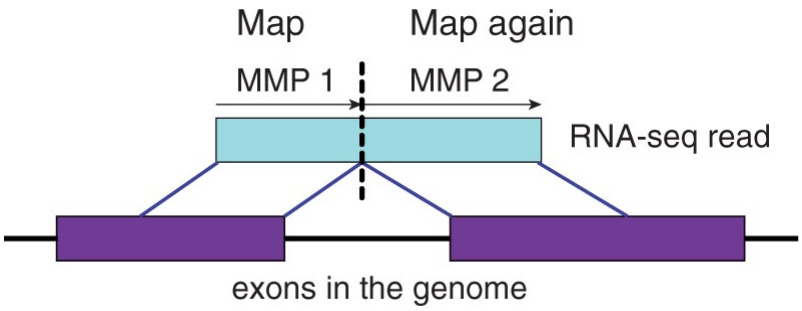

Fig. 5.2: Schematic representation of the MMP search in the STAR algorithm

5.2 UMI-tools

The UMI-tools deduplication algorithm [43] consists of three main steps. The first step before deduplication is extracting UMIs from raw reads to keep the sequence and remove the random nucleotides. The second and the most computationally intensive part is mapping reads. In our case, reads are mapped to the genome using STAR aligner described previously. After the reads are mapped, the final SAM file is converted to BAM file using samtools [56], set of utilities that manipulate alignments in the BAM format. In the next step, BAM file needs to be sorted and indexed in order to run the deduplication procedure on it. In our application, marked duplicates as well as all reads retained are needed. Therefore, the particular command was used to obtain final TSV file where each read is marked with its read group.

5.3 My_UMI_tool

As mentioned previously, the necessity for tools with low complexity leads us to design and develop a new algorithm shown in Fig. 5.3. The proposed tool has been developed in R programming language and is freely available on GitHub (`https://github.com/lujbarilikova/My_UMI_tool`). To provide accurate estimation of the total number of unique molecules, some publicly available tools were implemented in the algorithm as well. The presented tool avoids time consuming alignment before deduplication in order to design fast algorithm that can efficiently determine the absolute number of unique molecules by identifying duplicate reads.

To implement our method within the framework of removing PCR duplicates, we developed a command line tool called My_UMI_tool. Proposed method comprises the following stages: pre-processing of reads from the input file in FASTQ format, clustering reads by UMI, clustering reads using freely available VSEARCH tool [57], determination of starting gap count, correction of UMI errors and final identification of duplicates to generate the final FASTQ file with deduplicated reads as well as TSV file containing all reads, each of them marked with its read group. The above-mentioned stages are described in the following subsections.

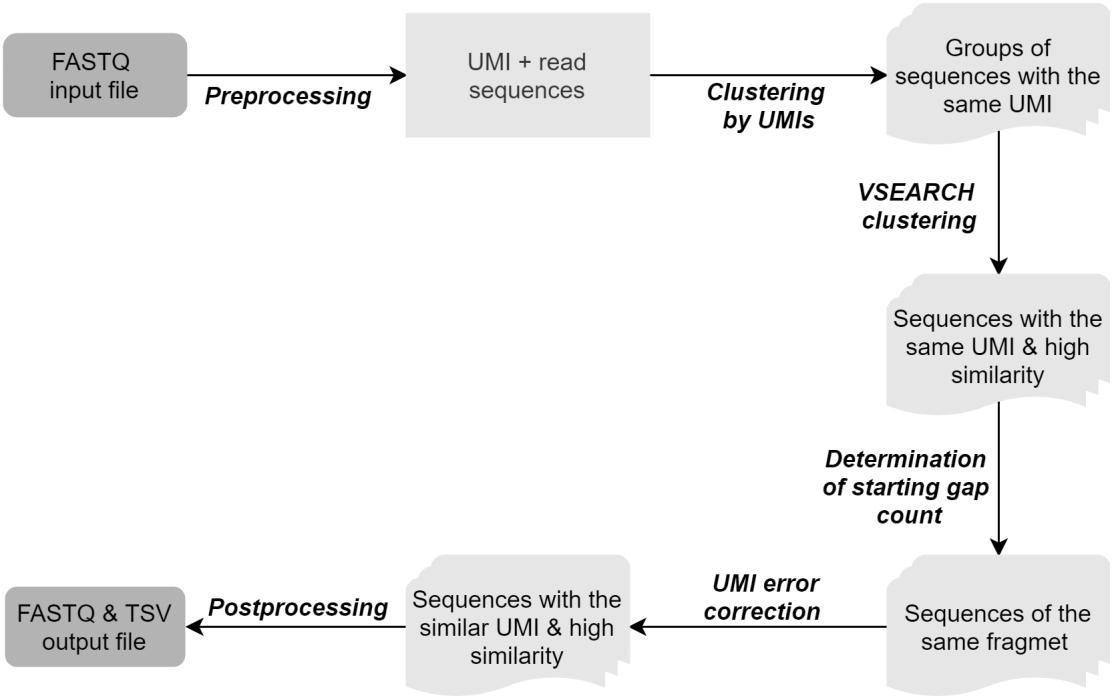

Fig. 5.3: Schematic diagram of My_UMI_tool algorithm

5.3.1 Pre-processing

To begin, the proposed pipeline requires FASTQ files as an input, which are text files comprising the nucleotide sequence and its quality (Phred) score for each base, represented as an ASCII character [58]. However, unprocessed FASTQ files cannot be used in downstream processing steps. Furthermore, UMI sequences, separated from the sequences of the fragments are stored in a header of each sequence in FASTQ file and need to be retained for further processes. Therefore, in the pre-processing step, shown in Fig. 5.4, UMIs are extracted from the header of each sequence in FASTQ file, added to the read name, and consequently stored with corresponding sequences to a separate data structure. After pre-processing, these UMIs are used to generate read groups consisting only of sequences with the similar UMIs. These steps are described in detail in the following sections respectively.

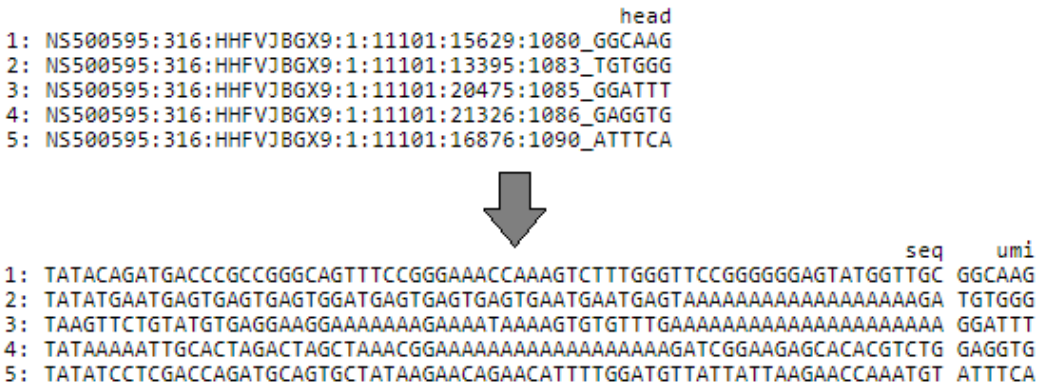

Fig. 5.4: Preprocessing step

5.3.2 Clustering by UMIs

In particular, no study, to our knowledge, has considered clustering approaches for recognizing UMIs that are expected to correspond to the same pre-amplified molecule as the first step when identifying PCR duplicates. In general, this step is based on the similarity between UMIs. As mentioned previously, UMI sequence is identified in a header of a each sequence in the FASTQ file, and then it is trimmed and each read sequence is annotated with the corresponding UMI. The main reason for handling UMIs first and afterwards the corresponding sequence of the read separately is that UMIs are usually much shorter than actual sequences. Most of the time, this approach is convenient, simple, and time-efficient.

As a result, sequences with the identical UMI are grouped together, suggesting that they may belong to the same fragment. The output of this step will be subsequently used as input to the next step, where groups of sequences with the same UMI will be clustered according to their similarity.

5.3.3 VSEARCH clustering

A second clustering step should be carried out on UMI clusters to further partition the sequences based on the non-UMI part of the reads. As indicated above, when merging multiple reads with the identical UMIs into a single cluster, checking that the rest of the sequence is also similar is recommended. The sequences within the cluster would be expected to differ only due to PCR and sequencing errors. Following this, reads of each group with the identical UMI are clustered by corresponding sequence similarity using VSEARCH, fast and accurate open source clustering tool used in a variety of bioinformatics applications.

In VSEARCH tool, de novo clustering is done using greedy and heuristic centroid-based algorithm, shown in Fig. 5.5, with a user-specified sequence similarity threshold. The algorithm works with initially empty database of centroid sequences. Each sequence from an input file is considered as query sequence and is subsequently clustered with the first centroid sequence with similarity threshold equal to or above the threshold.

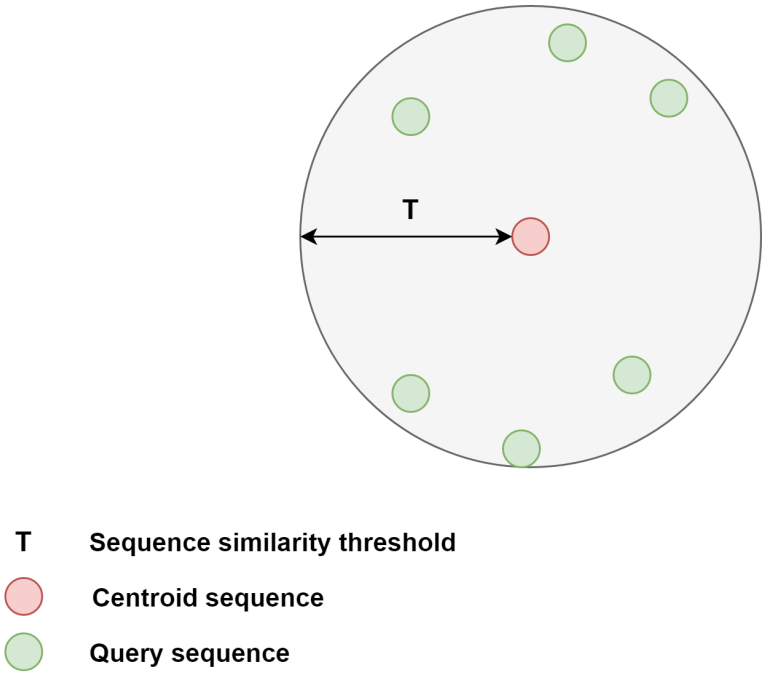

Fig. 5.5: Centroid-based algorithm

At first, sequences are quickly filtered according to statistics of shared words, which determines the similarity between sequences without the need to align them. In the next step, determining optimal global alignment between query sequence and the most promising candidates from database of centroid sequences in accordance with the number of words in common with the query, takes place. In other words, the alignment is firstly performed with the sequence having the largest number of words in common with the query sequence and then respectively with sequences with a decreasing number of shared words. If the query sequence is not clustered with the centroid sequence due to sequence similarity lower than the defined threshold, it becomes the centroid of a new cluster and is automatically added to the database.

Once the reads are partitioned into clusters, each corresponding to a single molecule, the next step is multiple sequence alignment using the center star method shown in Fig. 5.6, with the centroid as the center sequence, in order to build a consensus sequence for each cluster utilizing information from all reads in the corresponding cluster. To achieve this, all the pairwise alignments between the center sequence and the remaining sequences are merged. As a result, multiple alignment by adding sequences in decreasing order of similarity to center sequence is produced.

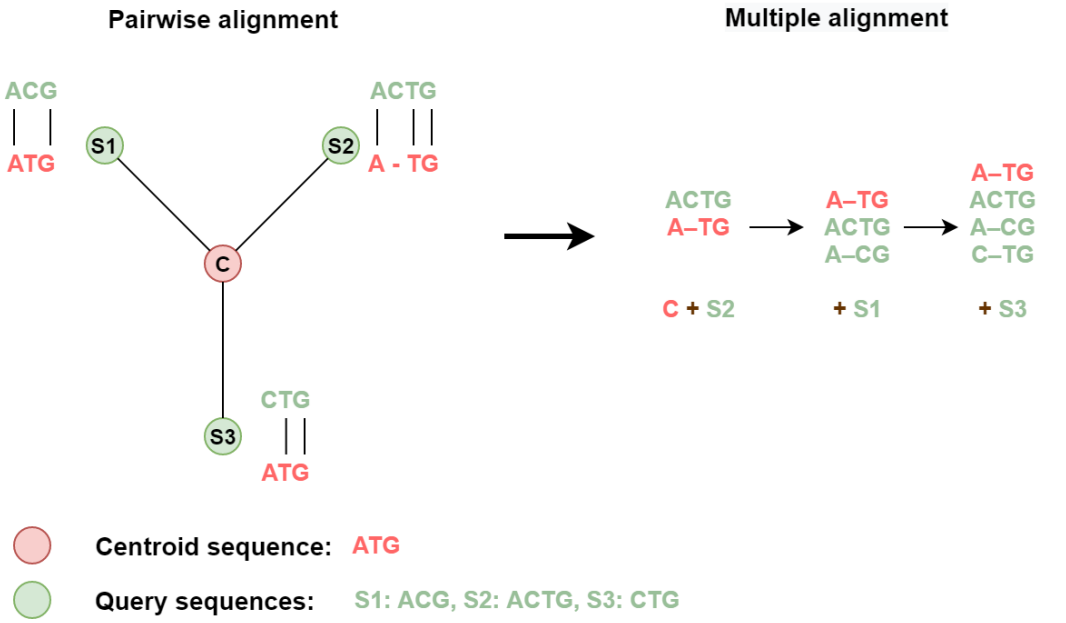

Fig. 5.6: Center star method for multiple sequence alignment

Determination of starting gap count

In addition, from generated multiple alignments, each sequence within-cluster is assigned with the number of starting gaps from the corresponding consensus sequence.

It is assumed that taking the number of gaps at the beginning of a consensus sequence into account could help to identify reads that originated from the same fragment. That is to say, each sequence is annotated with the corresponding UMI sequence, cluster number generated by the VSEARCH algorithm influenced by a given threshold, and finally, the number of starting gaps derived from a consensus sequence.

5.3.4 Correction of UMI errors

However, one of the main issues in our knowledge of UMIs is UMI error, which should be taken into consideration. One primary problem with UMI errors, resulting from nucleotide substitutions during PCR or nucleotide insertions or deletions during sequencing, is that additional artificial UMIs are created and therefore the estimation of the number of unique molecules might be negatively affected. In order to reduce the probability of wrong cluster allocation of reads and improve quantification using UMIs, nucleotide miscalling and substitution errors are not ignored.

Besides, as show in Fig. 5.7, there is also UMI collision depending mainly on the length of UMIs. Usually, the longer UMI length is, the higher diversity of UMIs is observed and therefore the number of UMIs is higher than the number of identical molecules. Unfortunately, this approach results in problems related to UMI errors when two UMIs become identical through amplification step by chance. Apart from that, it is also possible that two molecules are initially tagged with the same UMI. In addition, minimizing the impact of chimeric reads, which could also be an artifact of PCR amplification, is another important challenge.

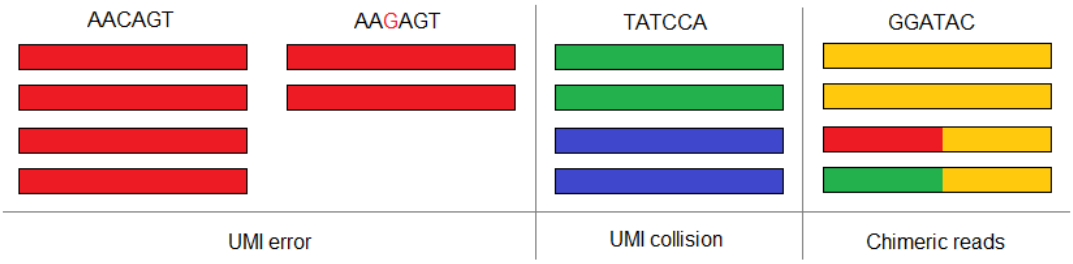

Fig. 5.7: UMI features, where each column represents reads with the same UMI, shown on the top of the column, and each color represents reads originating from the same molecule

There are many alternative methods available for solving these problems. One way to overcome this problem is to create clusters of sequences only with the identical UMIs, as described previously, and then examine these clusters. In our case, as shown in upper section of Fig. 5.8, all single-nucleotide different UMIs that could be observed for each unique combination of original UMI in the selected clusters are determined and eventually considered as similar. Two clusters of sequences can be clustered together only if their UMIs are similar. It is considered, that it will be sufficient for our purposes to deal only with the UMIs that are single-nucleotide different, and as a next step, only clusters whose counts of sequences are above a particular empirically chosen threshold are selected. Simply said, only those clusters are selected, whose counts of sequences are substantially higher than average counts. As shown in lower section of Fig. 5.8, these clusters are then considered as central clusters. Afterwards, remaining clusters assigned with the UMI similar to UMI of the central cluster, and at the same time, whose counts of sequences are substantially lower than average counts are selected and considered as nodes.

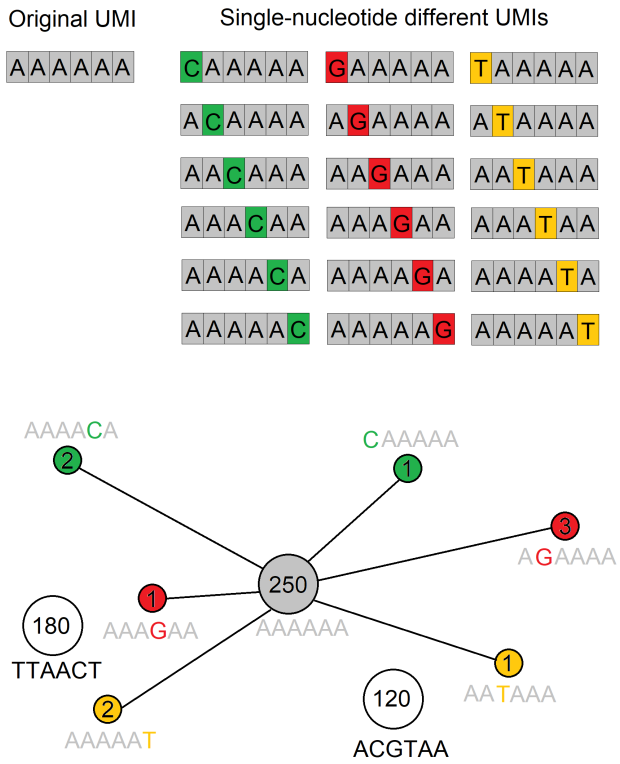

Fig. 5.8: My_UMI_tool method for resolving UMI errors (central cluster represented by grey color, node clusters represented by red, green and yellow with the number of sequences shown in the middle of the circle)

To finally cluster sequences with similar barcodes, the naive way is to compute pairwise string distances between read sequences corresponding to node clusters and read sequence from the central cluster. As a string metric, restricted Damerau-Levenshtein (DL) distance [59] is used to quantify the dissimilarity between two finite sequences. In theory, DL distance between two sequences is the minimum transform operations, such as insertions, deletions, substitutions or transpositions, required to change one sequence into the other. Accordingly, the lower the dissimilarity is, the closer the node clusters are to the central one. At the end, read sequences of node clusters with the dissimilarity lower than preferred threshold are assigned with the UMI corresponding to central cluster.

Typically, each UMI is observed multiple times and by this, we analyzed if UMIs originated from a single unique molecule prior to PCR amplification or from a combination of errors during PCR and sequencing or may originate from multiple unique molecules, which by chance have similar UMIs.

5.3.5 Post-processing

Once the information about UMI sequence, cluster number and the number of starting gaps is collected and assigned to the header of every single input sequence, desired output files can be readily produced by post-processing step.

The chosen parameters describing input sequences are used to identify reads with the same header as potential PCR duplicates, and remove them in order to generate final deduplicated FASTQ file. The FASTQ file output is subsequently used as an input for VSEARCH algorithm to generate BAM files described in a previous section.

To compare the results obtained using two different algorithms for deduplication, final results are also exported to a tab-separated file (TSV) containing all sequences from input FASTA file. Obviously, this file is not crucial for final deduplication but provides additional information about sequences in order to calculate summary statistics.

6 Application

The bench-marking pipeline is performed using Snakemake tool, known as a scalable bioinformatics workflow engine [60]. Snakemake workflows were designed to be human readable and are essentially Python based scripts defining rules that describes how to create output files from input files. The whole workflow works with dependencies between corresponding rules that arise from one rule that needs an input file, that is also an output file of another rule. To assess the performance of both tools under the impact of various conditions, the above-described pipeline is tested on simulated genomic data, and therefore, this section also contains a summary of the method used to simulate data from chosen reference sequence.

6.1 Data simulation

The purpose of simulation shown in Fig. 6.1 is to generate synthetic next-generation reads for which original UMI is known. This step is essential for testing proposed bench-marking pipeline to compare My_UMI_tool performance against UMI-tools.

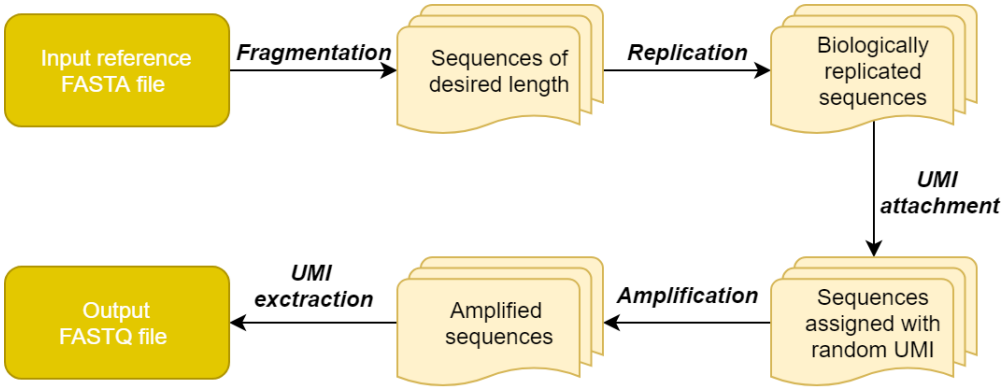

Fig. 6.1: Schematic diagram of data simulation

Suggested simulation can be seen as two stage process: generation of biological duplication and generation of technical duplication. At first, target sequences from required input reference FASTA file are fragmented to a desired length using sliding window size of 75 nucleotides with a step size of 1 nucleotide. Obtained sequences are randomly replicated to simulate biological duplication. Afterwards, to each individual read sequence, random UMI sequence is attached. The UMI sequence needs to be preserved, and therefore, it is also assigned to the head of a corresponding read sequence. These sequences are then exported to FASTA file, as the next step only accepts input sequences in FASTA format.

Technical duplication or library amplification in the second step is performed using ART, a next-generation sequencing read simulator [61]. This tool simulates sequencing reads by mimicking real sequencing process with empirical error models or quality profiles summarized from large re-calibrated sequencing data. To generate final FASTQ file, simulation of Illumina sequencers is used and since this technology reads out one base at a time, the main error mode is substitution rather than insertion or deletion. At the end, UMIs are extracted from the read sequences and assigned to the head of the read as well. Final head of each read sequence then consist of the original UMI sequence and UMI sequence after the amplification, so the UMI errors can be observed.

6.2 Results and discussion

Tools proposed in the pipeline are validated with both simulated and experimental datasets. The advantage of a simulated dataset is that it is allowed to assess performance where the number of duplicated, as well as unique reads are known and can be afterwards used as an objective measure of performance. On the other hand, experimental dataset provides the opportunity to evaluate whether the results lead to biologically relevant conclusions. The detailed information about simulated datasets are shown in Tab. 6.1. Six types of simulated datasets containing different number of replicated and amplified sequences are examined. As a reference FASTA file, BRNO-ONCO (BRONCO) panel provided by CEITEC-MU, was used. The BRONCO panel, containing 296 genes, is an attempt to reveal germinal pathogenic variants in genes considered as genetic risk component of a tumour disease. The sequencing was performed on an Illumina NextSeq 500 machine with sequencing library prepared using the SureSelect HS XT technology.

Tab. 6.1: Detailed information about simulated datasets

	Dataset1	**Dataset2**	**Dataset3**	**Dataset4**	**Dataset5**	**Dataset6**
Number of sequences from reference FASTA	3200	1420	16 347	890	212	1420
Random replication (from:to)	1:50	1:50	1:50	1:10	1:3	1:10
Number of replicated sequences	82 466	36 025	416 457	5 029	639	8 014
Number of reads per amplicon	20	40	5	1000	10 000	500
Number of sequences in final FASTQ file	1 649 320	1 441 000	2 082 285	5 029 000	5 112 000	4 014 000

6.2.1 Results from simulated datasets

For each dataset, two by two contingency table summarising the results from both tools are constructed. Multiple contingency tables are then used to determine sensitivity, specificity, accuracy, precision and F1-score as follows:

$$Sensitivity(Recall) = TP/(TP + FN), \tag{6.1}$$

$$Specificity = TN/(TN + FP), \tag{6.2}$$

$$Accuracy = (TP + TN)/(TP + FP + FN + TN), \tag{6.3}$$

$$Precision = TP/(TP + FP), \tag{6.4}$$

$$F1 - score = 2 \cdot (Recall \cdot Precision)/(Recall + Precision). \tag{6.5}$$

The reporting statistical measures are defined using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) counts where:

- TP represents duplicate read marked as duplicate,
- TN represents unique read marked as unique,
- FP represents unique read marked duplicate,
- FN represents duplicate read marked as unique.

Sensitivity, also known as recall or true positive test, is defined as the proportion of duplicate reads that are marked as duplicates. In other words, a highly sensitive tool is one that correctly identifies duplicate reads. Specificity, on the other hand, evaluates the ability of a tool to determine the unique reads correctly or to determine reads that are not duplicates. In general, if the sensitivity is high, specificity is usually relatively low. It means that a tool is good at determining which one are duplicate reads, but it also means that tool has a fairly high rate of false positives. Likewise, high specificity means that the tool has lower sensitivity and quite high rate of false negatives. Accuracy is simply a ratio of correctly predicted observation to the total observations and estimates how correct a tool differentiates duplicate and unique reads or how close a decision if the read is duplicate or not is to its true state. Precision measures how many reads marked as duplicates are actual duplicates, i.e. the percentage of correct predictions. Precisely working tool also means how repeatable is its measurement. F1-score is the harmonic average of the precision and recall and takes both, false positives and false negatives into account. In contrast to accuracy, F1-score should give a better measure of the incorrectly classified reads as the accuracy takes only true positives and true negatives into consideration.

The first dataset consists of 1 649 320 read sequences and only 82 466 of them are truly unique. As seen in Tab. 6.2, My_UMI_tool marked 83 395 of the read sequences as unique and 81 398 of them were identified correctly. This gave the tool a specificity of 98,70 %, which describes its ability to correctly determine if the read is unique or not. Accordingly, from all the 1 566 854 duplicate read sequences, the tool marked 1 564 857 of them correctly as duplicates, which gave the tool a sensitivity of 99,87 %. From all the observations, 99,81 % were identified correctly, which defines an accuracy of a tool. From all the 1 565 925 read sequences marked as duplicates, 1 564 857 are actual duplicates, which consequently resulted in the precision of 99,93 %. As seen in 6.3, similar results were obtained using UMI-tools. However, besides sensitivity, according to the statistical measures, the performance of My_UMI_tool is either better or the same.

Tab. 6.2: My_UMI_tool Dataset1

Dataset1	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	1 564 857	1 068	1 565 925
Marked as Unique	1 997	81 398	83 395
Column total	1 566 854	82 466	1 649 320

Tab. 6.3: UMI-tools Dataset1

Dataset1	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	1 545 301	1 641	1 546 942
Marked as Unique	1 390	80 284	81 674
Column total	1 546 691	81 925	1 628 616

The second dataset is quite similar to the first one, except that the number of replicated sequences is reduced from 82 466 to 36 025, as well as the number of reads per amplicon is increased from 20 to 40. Essentially, 36 025 is the number of unique sequences that needs to be reached. As seen in Tab. 6.4 and Tab. 6.5, by My_UMI_tool 35 203 of them were determined as unique, whereas by UMI-tools only 34 590 of them were determined as unique. Despite sensitivity, as in the previous case, My_UMI_tool performs the same or even better than UMI-tools when processing this dataset.

Tab. 6.4: My_UMI_tool Dataset2

Dataset2	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	1 403 339	822	1 404 161
Marked as Unique	1 636	35 203	36 839
Column total	1 404 975	36 025	1 441 000

Tab. 6.5: UMI-tools Dataset2

Dataset2	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	1 382 040	1 312	1 383 352
Marked as Unique	1 249	34 590	35 839
Column total	1 383 289	35 902	1 419 191

The third dataset consists of 416 457 unique read sequences and the number of reads per amplicon is set only to 5. Consequently, in this dataset, biological duplication predominates. The total number of read sequences obtained in the final FASTQ file is 2 082 285. From the final results shown in Tab. 6.14 and Tab. 6.15, it can be seen that the best specificity is obtained compared to other results from the remaining datasets. Despite the satisfactory values of specificity of 99,33 % and 99,06 % respectively for My_UMI_tool and UMI-tools, analysing this dataset and its results from Tab. 6.6 and Tab. 6.7, the precision, accuracy, F-score, and sensitivity produce the lowest performance profiles when comparing to other datasets.

Tab. 6.6: My_UMI_tool Dataset3

Dataset3	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	1 658 229	2 777	1 661 006
Marked as Unique	7 599	413 680	421 279
Column total	1 665 828	416 457	2 082 285

Tab. 6.7: UMI-tools Dataset3

Dataset3	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	1 642 746	3 855	1 646 601
Marked as Unique	2 112	408 291	410 403
Column total	1 644 858	412 146	2 057 004

The fourth dataset, containing 5 029 000 read sequences, was simulated with significantly higher number of reads per amplicon, as shown in Tab. 6.1. Therefore, it is assumed that in this dataset, technical duplication plays a predominant role when compared to previous simulations. As seen in Tab. 6.8, by M_UMI_tool, from all the 5 020 734 read sequences marked as duplicates, only 227 read sequences are false positive, which consequently resulted in the precision of 100 %. In the case of UMI-tools with the results in Tab. 6.9 and according to Tab. 6.15, the specificity reached only 87,99 %, which is substantially less than obtained by M_UMI_tool, where specificity of 95,49 % was achieved. However, from all the datasets, the corresponding specificity is the lowest obtained by both tools.

Tab. 6.8: My_UMI_tool Dataset4

Dataset4	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	5 020 507	227	5 020 734
Marked as Unique	3 464	4 802	8 266
Column total	5 023 971	5 029	5 029 000

Tab. 6.9: UMI-tools Dataset4

Dataset4	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	4 891 873	604	4 892 477
Marked as Unique	749	4 425	5 174
Column total	4 892 622	5 029	4 897 651

The fifth dataset is quite similar to the previous one, however, due to very low number of replicated sequences and at the same time, a very high number of reads per amplicon, technical duplication has a much greater effect on the final dataset. Results shown in Tab. 6.10 and Tab. 6.11 are, as expected, also very similar to those observed in the previous dataset. As the technical duplication increases, the capability to determine the unique reads correctly decreases and, therefore, My_UMI_tool shows 96,24 % of specificity with a sensitivity of 99,96 %, while UMI-tool shows 91,71 % of specificity with a sensitivity of 99,99 %. Consequently, besides specificity, remaining statistical measures reached the best results over all the datasets decisively.

Tab. 6.10: My_UMI_tool Dataset5

Dataset5	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	5 109 084	24	5 109 108
Marked as Unique	2277	615	2 892
Column total	5 111 361	639	5 112 000

Tab. 6.11: UMI-tools Dataset5

Dataset5	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	5 078 300	53	5 078 353
Marked as Unique	115	586	701
Column total	5 078 415	639	5 079 054

The last dataset that is considered represents a combination of the biological and technical duplication. As the technical duplication slightly prevails, the results are similar and as shown in Tab. 6.14 and Tab. 6.15, and besides specificity, all the statistical measures by both tools achieved nearly 100 %. Accordingly, My_UMI_tool shows 96,84 % of specificity, while UMI-tool shows only 91,75 % of specificity. From all the read sequences marked by My_UMI_tool as duplicates, only 254 of them were marked incorrectly. In case of UMI-tools, there were 662 read sequences incorrectly marked as duplicates. Additionally, from Tab. 6.12 and Tab. 6.13, it may look like there are a lot of false negatives in the case of My_UMI_tool, but the results clearly indicates the much bigger difference in true positives between these two tools.

Tab. 6.12: My_UMI_tool Dataset6

Dataset6	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	4 003 095	254	4 003 349
Marked as Unique	2 877	7 774	10 651
Column total	4 005 972	8 028	4 014 000

Tab. 6.13: UMI-tools Dataset6

Dataset6	**Duplicate**	**Unique**	**Row total**
Marked as Duplicate	3 908 129	662	3 908 791
Marked as Unique	839	7 366	8 205
Column total	3 908 968	8 028	3 916 996

The final statistics for six simulated datasets used for testing are shown in Tab. 6.14 and Tab. 6.15. In general, the higher the number of reads per amplicon is set, the better the sensitivity and at the same time, the worse the specificity is obtained. However, there were no significant differences between My_UMI_tool and UMI-tools as far as the accuracy is concerned. The single most striking observation to emerge from the comparison of the tools was the specificity of 97,39 % achieved by My_UMI_tool and specificity of 94,14 % achieved by UMI-tools. What is important to mention is the fact that UMI-tools allow multi-mapping reads that are usually removed and, therefore, as seen from contingency tables, the total number of the read sequences marked either as a duplicate or unique differ from the original number of read sequences. On the other hand, as seen from Tab. 6.14, the time performance was slightly disappointing. This was probably as a result of repeatedly writing the results to files and reloading them in an effort to process the data, but the trade-off between longer computation times before alignment associated with larger datasets and better classification performance is usually worthwhile.

Tab. 6.14: Final statistics for My_umi_tool

	Sensitivity [%]	**Specificity** [%]	**Accuracy** [%]	**Precision** [%]	**F1-score** [%]	**Run-time** [s]
Dataset1	99,87	98,70	99,81	99,93	99,90	2258
Dataset2	99,88	97,72	99,83	99,94	99,91	9198
Dataset3	99,54	99,33	99,50	99,83	99,69	2256
Dataset4	99,93	95,49	99,93	100,0	99,96	1460
Dataset5	99,96	96,24	99,95	100,0	99,98	2253
Dataset6	99,93	96,84	99,92	99,99	99,96	2563
Average	99,85	97,39	99,82	99,95	99,90	3331

Tab. 6.15: Final statistics for UMI-tools

	Sensitivity [%]	**Specificity** [%]	**Accuracy** [%]	**Precision** [%]	**F1-score** [%]	**Run-time** [s]
Dataset1	99,91	98,00	99,81	99,89	99,90	420
Dataset2	99,91	96,35	99,82	99,91	99,91	208
Dataset3	99,87	99,06	99,71	99,77	99,82	421
Dataset4	99,98	87,99	99,97	99,99	99,99	129
Dataset5	99,99	91,71	100,0	100,0	100,0	150
Dataset6	99,98	91,75	99,96	99,98	99,98	334
Average	99,94	94,14	99,92	99,92	99,93	277

6.2.2 Comparison with real datasets

In order to evaluate the performance of My_UMI_tool tool on real data, samples prepared by two different protocols, Formalin-Fixation and Paraffin-Embedding (FFPE) and Freshly Frozen (FF), from Chronic Lymphocytic Leukemia (CLL) patients are studied. Library of sequences was subsequently generated by all-in-one library preparation protocol QuantSeq 3' mRNA-Seq Library Prep Kit with an additional module with UMIs [62].

As shown in Tab. 6.16, when comparing results from the proposed My_UMI_tool tool method to those obtained by available tool UMI-tools for handling UMIs in NGS data sets, it must be pointed out that a high percentage of sequences in both samples are clustered and mapped identically in both tools. From this standpoint, it can be considered that these sequences are mapped correctly.

In this samples, UMIs are six bases long and it is important to highlight the fact that even UMIs with three error bases were grouped by UMI-tools together and this could be considered as very exaggerative. In line with the ideas of UMI-tools and its acceptance of UMI errors, it can be concluded that 5.76 % of sequences in FFPE sample and 14.67 % of sequences in FF sample are therefore clustered by My_UMI_tool more complexly in smaller groups.

In contrast, My_UMI_tool expects UMIs to be only with one error base at maximum and corresponding 1,83 % of sequences in FFPE sample and 1.5 % of sequences in FF sample, more complexly grouped by UMI-tools, could be the result of mentioned alignment before deduplication. After the alignment, reads aligned to the genome with the same mapping position are grouped together and then, by examining all UMIs at the single locus, clustered by different methods to resolve UMI errors.

Tab. 6.16: Final statistics

Sample	**FFPE sample** [%]	**FF sample** [%]
Clustered in both tools	92,41	83,83
Clustered in My_UMI_tool	5,76	14,67
Clustered in UMI-tools	1,83	1,50

7 Conclusions

As stated in the Introduction, the main purpose of this work was to design an algorithm that can efficiently determine the absolute number of unique molecules by identifying duplicate reads in an input file. In general, the presented tools to solve PCR errors usually start with time-consuming alignment before deduplication. Moreover, multi-mapping reads defined as sequences that map more than once on the genome due to multiple copies of a gene, are typically allowed. This makes it difficult to distinguish between genuinely multi-mapping reads and reads that just come from multiple fragments of the same gene. Many tools ignore these sequences as defaults, which means that at least 20-30% of the data are lost.

The design of My_UMI_tool was based on a comprehensive study of the strengths of each available tool, where the reader can look up the individual tool in the fourth section of this work. In addition, these findings provide additional information about all of the disadvantages of the mentioned tools and why there is such interest to continually innovate and develop new tools. Proposed method comprises the following stages: pre-processing of reads from the input file in FASTQ format, clustering reads by UMI, clustering reads with the same UMI according their similarity, determination of starting gap count, correction of UMI errors and final identification of duplicates to generate the final FASTQ file with deduplicated reads as well as TSV file containing all reads, each of them marked with its read group. Additionally, to evaluate the performance of My_UMI_tool under the impact of various conditions, the above-described algorithm is tested on simulated genomic data, as well as experimental data. The performance is compared with the UMI-tools as it is one of the most commonly used tools with high prediction accuracy. The results show that avoiding time-consuming alignment before deduplication does not seem to impact the final determination of the absolute number of unique molecules and are equal to or better than results that are currently accepted.

In summary, this work argued that My_UMI_tool is a valuable tool for deduplicating next-generation sequencing data using UMIs, where duplicate reads are removed from the sample to prepare data for downstream analysis. Apart from existing tools, My_UMI_tool is designed to avoid alignment before deduplication and, therefore, will fill the gap in the currently available tools. Although time performance is not ideal, it is still believed that this tool will be useful in applications such as analysis of transposable elements or Alu elements, which make up more than 10% of the human genome. From this point of view, getting a set of correctly deduplicated reads before an alignment is crucial and, therefore, will significantly solve the problem with multi-mapping reads.

Bibliography

[1] TRINGE, Susannah Green; RUBIN, Edward M. Metagenomics: DNA sequencing of environmental samples. *Nature reviews genetics*, 2005, 6.11: 805–814.

[2] WHEELER, David A., et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 2008, 452.7189: 872–876.

[3] RABBANI, Bahareh, et al. Next generation sequencing: implications in personalized medicine and pharmacogenomics. *Molecular BioSystems*, 2016, 12.6: 1818–1830.

[4] LUPSKI, James R., et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *New England Journal of Medicine*, 2010, 362.13: 1181–1191.

[5] ILLUMINA. *An introduction to next-generation sequencing technology* [online]. Illumina, 2017, Pub.No. 770-2012-008-B. Available from: `https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`

[6] HUGHES, Christopher; MA, Bin; LAJOIE, Gilles A. De novo sequencing methods in proteomics. In: *Proteome Bioinformatics*. Humana Press, 2010. p. 105–121.

[7] ALTELAAR, AF Maarten; MUNOZ, Javier; HECK, Albert JR. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 2013, 14.1: 35–48.

[8] JOVEL, Juan, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology*, 2016, 7: 459.

[9] WOOD, Derrick E.; SALZBERG, Steven L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 2014, 15.3: R46.

[10] HUTTENHOWER, Curtis, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 2012, 486.7402: 207.

[11] COHEN, Jonathan; POWDERLY, William G; OPAL Steven M. *Infectious Diseases*. 4. Elsevier Health Sciences, 2016. ISBN 978-0702063381

[12] GAWAD, Charles; KOH, Winston; QUAKE, Stephen R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 2016, 17.3: 175.

[13] CHEN, Geng; SHI, Tieliu. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in genetics*, 2019, 10: 317.

[14] WANG, Jian; SONG, Yuanlin. Single cell sequencing: a distinct new field. *Clinical and translational medicine*, 2017, 6.1: 10.

[15] PAPALEXI, Efthymia; SATIJA, Rahul. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 2018, 18.1: 35.

[16] PETTERSSON, Erik; LUNDEBERG, Joakim; AHMADIAN, Afshin. Generations of sequencing technologies. *Genomics*, 2009, 93.2: 105–111.

[17] HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 2016, 107.1: 1–8.

[18] KCHOUK, Mehdi; GIBRAT, Jean-François; ELLOUMI, Mourad. Generations of sequencing technologies: from first to next generation.*Biology and Medicine*, 2017, 9.3.

[19] WARNER Patrick, et al. *Sanger sequencing White paper* [online]. University of Minnesota Genomics Center. Available from: `http://genomics.umn.edu/downloads/sanger_white_paper.pdf`

[20] EL-METWALLY, Sara; OUDA, Osama M.; HELMY, Mohamed. *Next generation sequencing technologies and challenges in sequence assembly*. Springer Science & Business, 2014.

[21] MELDRUM, Cliff; DOYLE, Maria A.; TOTHILL, Richard W. Next-generation sequencing for cancer diagnostics: a practical perspective. *The Clinical Biochemist Reviews*, 2011, 32.4: 177.

[22] MOORTHIE, Sowmiya; MATTOCKS, Christopher J.; WRIGHT, Caroline F. Review of massively parallel DNA sequencing technologies. *The HUGO journal*, 2011, 5.1-4: 1–12.

[23] KULSKI, Jerzy K. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. *Next Generation Sequencing–Advances, Applications and Challenges*, 2016, 3–60.

[24] AMBARDAR, Sheetal, et al. High throughput sequencing: an overview of sequencing chemistry. *Indian journal of microbiology*, 2016, 56.4: 394–404.

[25] ESCALANTE, Ana E., et al. The study of biodiversity in the era of massive sequencing. *Revista Mexicana de Biodiversidad*, 2014, 85.4: 1249–1264.

[26] KUMAR, Santosh; BANKS, Travis W.; CLOUTIER, Sylvie. SNP discovery through next-generation sequencing and its applications. *International journal of plant genomics*, 2012, 2012.

[27] VOELKERDING, Karl V.; DAMES, Shale A.; DURTSCHI, Jacob D. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 2009, 55.4: 641–658.

[28] SCHADT, Eric E.; TURNER, Steve; KASARSKIS, Andrew. A window into third-generation sequencing. *Human molecular genetics*, 2010, 19.R2: R227–R240.

[29] RHOADS, Anthony; AU, Kin Fai. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 2015, 13.5: 278–289.

[30] EISENSTEIN, Michael. The battle for sequencing supremacy. *Nature biotechnology*, 2012, 30.11: 1023.

[31] GOODWIN, Sara; MCPHERSON, John D.; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 2016, 17.6: 333.

[32] LU, Hengyun; GIORDANO, Francesca; NING, Zemin. Oxford Nanopore MinION sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, 2016, 14.5: 265–279.

[33] JAIN, Miten, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, 2016, 17.1: 239.

[34] FRIEDMANN, Theodore, et al. *Advances in genetics*. Academic Press, 1996.

[35] BLOW, Nathan. DNA sequencing: generation next-next. *Nature Methods*, 2008, 5.3: 267–274.

[36] STARK, Rory; GRZELAK, Marta; HADFIELD, James. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 2019, 20.11: 631–656.

[37] KUMAR, Ravi, et al. A high-throughput method for Illumina RNA-Seq library preparation. *Frontiers in plant science*, 2012, 3: 202.

[38] OIKONOMOPOULOS, Spyros, et al. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Scientific reports*, 2016, 6.1: 1–13.

[39] GONZALEZ-GARAY, Manuel L. Introduction to isoform sequencing using pacific biosciences technology (Iso-Seq). In: *Transcriptomics and Gene Regulation*. Springer, Dordrecht, 2016. p. 141–160.

[40] GARALDE, Daniel R., et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods*, 2018, 15.3: 201.

[41] DOYLE Maria; PHIPSON Belinda; DASHNOW Harriet. *RNA-Seq reads to counts (Galaxy Training Materials)* [online], 2020. Available from: `http://genomics.umn.edu/downloads/sanger_white_paper.pdf`

[42] FU, Yu, et al. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *Bmc Genomics*, 2018, 19.1: 531.

[43] SMITH, Tom; HEGER, Andreas; SUDBERY, Ian. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, 2017, 27.3: 491–499.

[44] ISLAM, Saiful, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 2014, 11.2: 163.

[45] CLEMENT, Kendell, et al. AmpUMI: design and analysis of unique molecular identifiers for deep amplicon sequencing. *Bioinformatics*, 2018, 34.13: i202–i210.

[46] SENA, Johnny A., et al. Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Scientific reports*, 2018, 8.1: 1–13.

[47] CHEN, Shifu, et al. gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. *BMC bioinformatics*, 2019, 20.23: 606.

[48] BROAD INSTITUTE. *Picard toolkit* [online]. GitHub repository, 2019. Available from: `http://broadinstitute.github.io/picard/`

[49] MANGUL, Serghei, et al. UMI-Reducer: Collapsing duplicate sequencing reads via Unique Molecular Identifiers. *bioRxiv*, 2017, 103267.

[50] PAREKH, Swati, et al. zUMIs-a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*, 2018, 7.6: giy059.

[51] DOBIN, Alexander, et al. STAR: ultrafast universal RNA-seq aligner.*Bioinformatics*, 2013, 29.1: 15–21.

[52] LIAO, Yang; SMYTH, Gordon K.; SHI, Wei. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 2014, 30.7: 923–930.

[53] FENNELL Tim; HOMER Nails. *fgbio toolkit* [online]. Fulcrum Genomics, 2019. Available from: `http://fulcrumgenomics.github.io/fgbio/`

[54] GIRARDOT, Charles, et al. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC bioinformatics*, 2016, 17.1: 419.

[55] SMITH, Andrew M., et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic acids research*, 2010, 38.13: e142–e142.

[56] LI, Heng, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, 25.16: 2078–2079.

[57] ROGNES, Torbjørn, et al. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 2016, 4: e2584.

[58] COCK, Peter JA, et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 2010, 38.6: 1767–1771.

[59] HOSANGADI, Sandeep. Distance measures for sequences. *arXiv preprint arXiv:1208.5713*, 2012.

[60] KÖSTER, Johannes; RAHMANN, Sven. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 2012, 28.19: 2520–2522.

[61] HUANG, Weichun, et al. ART: a next-generation sequencing read simulator. *Bioinformatics*, 2012, 28.4: 593–594.

[62] MOLL, Pamela, et al. QuantSeq 3' mRNA sequencing for RNA quantification. *Nature methods*, 2014, 11.12: 972.

List of abbreviations

ASCII	American Standard Code for Information Interchange
ATP	Adenosine Phospho-Sulphate
BAM	Compressed binary version of a SAM
BC	Barcode
CCD	Charge-Coupled Device
cDNA	Complementary DNA
DL	Damerau-Levenshtein
DNA	Deoxyribonucleic Acid
dsDNA	Double-Stranded DNA
FF	Freshly Frozen
FFPE	Formalin-Fixation and Paraffin-Embedding
QC	Quality Control
mtDNA	Mitochondrial DNA
NGS	Next Generation Sequencing
ONT	Oxford-Nanopore Technology
PCR	Polymerase Chain Reaction
rRNA	ribosomal Ribonucleic Acid
RNA	Ribonucleic Acid
RNA-seq	RNA-sequencing
UMI	Unique Molecular Identifier
SAM	Sequence Alignment Map
SGS	Second Generation of Sequencing
SMRT	Single-Molecule, Real-Time
TSV	Tab-Separated Values
ZMW	Zero Mode Waveguide