Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Characterisation of miRNA variation in Small RNA-Seq data

Frida Moi
Civil Engineering, Chemistry and Biotechnology - Bioinformatics

Abstract

Evidence suggests that miRNA signatures could have clinical applications as biomarkers for diagnostic and prognostic purposes. In standard analyses, miRNAs are considered to exist as well-defined features with a precise start and stop positions. In reality, they exist as a population of similar but slightly different isoforms - isomiRs. The project has developed methods to investigate isomiR populations and studied their variation between healthy and cancer patients. The methods presented here have been implemented in an analysis pipeline that allows standardised and scalable analysis of raw NGS datasets. The value of this approach has been demonstrated by investigating publicly available datasets to identify statistically significant changes in isomiR populations in various cancers.

Sammendrag

Forskning tilsier at miRNA-signaturer kan ha kliniske applikasjoner i diagnostisk- og prognostisk sammenheng. I konvensjonelle analyser antas miRNA å være veldefinerte enheter med nøyaktige start- og slutt-posisjoner. I virkeligheten eksisterer de som en variert populasjon av lignende, men noe ulike, isomerer. Gjennom dette prosjektet har det blitt utviklet metoder for å undersøke hvordan miRNA-populasjoner varierer mellom friske og syke. Metodene har blitt implementert i eksisterende programvare som muliggjør standardisert og skalerbar analyse av sekvenseringsdata. Verdien av denne tilnærmingen har blitt demonstrert ved å analysere offentlig tilgjengelige datasett for å identifisere statistisk signifikante endringer i miRNA-populasjoner i ulike krefttyper.

Acknowledgements

This thesis was written as part of my master's program in Civil Engineering at NMBU, the Faculty of Chemistry and Biotechnology, in collaboration with the Computational Biology Group at Oslo University Hospital.

First and foremost, I want to share my deepest appreciation for my supervisors.

> Professor Simon Rayner at OUS, for providing continuous support, advice, and encouragement, with an impeccable blend of insight and humour.
>
> Associate professor Jon Olav Vik at NMBU, for guiding the way with uplifting talks and invaluable feedback.
>
> Dr. Pavel Vazquez Faci at OUS, for always lending a helping hand and spreading positivity.

I also want to extend my sincere thanks to the Computational Biology Group at OUS for the warm welcome, the kindness, and the laughs. I am grateful to have been able to work with all of you.

Also, thank you,

> To the Bioinformatics and Statistics group at NMBU.
> To the Waszak group at the Norwegian Centre for Molecular Medicine.
>
> You have inspired me to follow this path.

And finally, thank you to Friends & Family for always supporting me.

List of Abbreviations

ATAC-Seq	Assay for transposable-accessible chromatin sequencing
CBGOUS	Computational Biology Group at Oslo University Hospital
ChIP-Seq	Chromatin immunoprecipitation sequencing
CPM	counts per million
CRC	colorectal cancer
GBM	glioblastoma
GWAS	genome-wide association studies
LAC	lung adenocarcinoma
miRISC	microRNA induced silencing complex
miRNA	microRNA
MMRN	microRNA-mRNA regulatory network
mRNA	messenger RNA
ncRNA	non-coding RNA
NGS	next generation sequencing
nt	nucleotide
PC	prostate cancer
pre-miRNA	precursor microRNA
pri-miRNA	primary microRNA
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SRA	sequence read archive
UTR	untranslated region

Contents

List of figures

List of tables

1 Introduction

1.1 The genetic code

The Central Dogma of molecular biology presented by Francis Crick in 1958 [1], stated that genetic information is carried from DNA to RNA to protein [2]. To this day it remains a fundamental backbone of genomics.

Genes, in the traditional sense, are regions of the genomic sequence that produce functional proteins. The DNA sequence is first copied, producing messenger RNAs (mRNAs), through a process called transcription. mRNAs are in turn translated into proteins, with each triplet of nucleotides corresponding to an amino acid. Not all parts of the mRNA are effectively translated; both ends have untranslated regions (UTRs), and the mRNA often has regions of non-coding sequences called introns (Figure 1). The coding regions are commonly called exons.

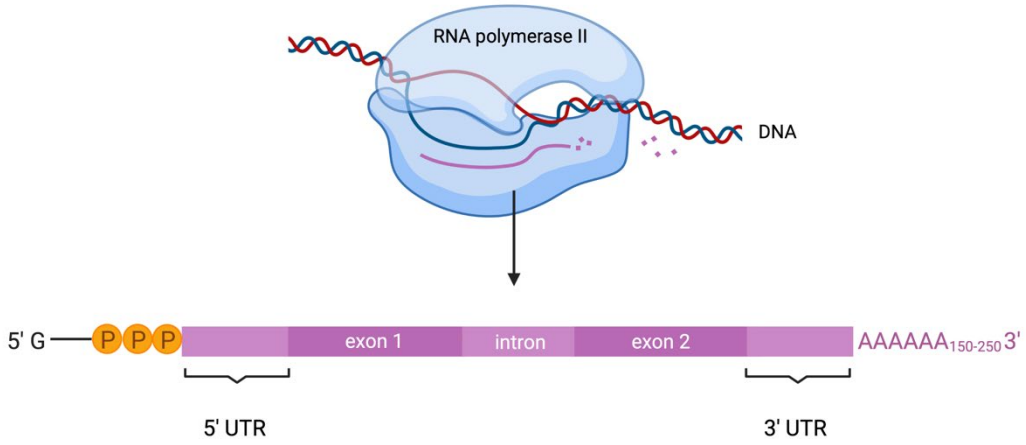


Figure 1. An mRNA transcribed by RNA polymerase II. The mRNA consists of coding sequences (exons), non-coding sequences (introns), and an untranslated region on either end. In addition, post-translational modifications are made to 5' (capping) and 3' (polyadenylation) ends [3].

It was long believed that most of our genome consisted of protein coding genes. Thus, one of the most surprising discoveries of modern biology was the finding that no more than 2% of the human genome is in fact protein coding. The first draft of the human genome was made available in 2001 by means of the Human Genome Project [4]. Since then, millions of genomes have been sequenced, which has uncovered a wealth of genetic variation.

Variation in the human genome is often in the form of single nucleotide variants (SNVs). SNVs encompass any single nucleotide changes in the genomic sequence, whether germline or somatic, common, or rare. SNVs are generally called single nucleotide polymorphisms (SNPs) if they are germline variants naturally occurring in at least 1% of the population. These terms are similar and are sometimes used interchangeably, but they have different definitions.

SNPs are widespread throughout the genome, and most are found outside of coding regions. As many as 43% of trait-associated SNPs have been reported to be intergenic, and 45% to be intronic [5]. Traits in this context are qualitative or quantitative characteristics of an individual such as hair colour, disease (qualitative) or height (quantitative).

Since the completion of The Human Genome Project, there have been tremendous breakthroughs in the field of medical genetics, linking genotypes to phenotypes and uncovering disease causing- and associated genetic variants.

Genome-wide association studies (GWAS) have been instrumental in these findings. GWAS are mainly used to identify SNPs that are associated with traits and disease. GWAS SNP arrays have unveiled the genetic origins of many diseases, especially Mendelian disorders such as cystic fibrosis [6]. However, for highly complex malignancies like cancer, the associations are less obvious.

While GWAS are indeed based on whole genome sequencing, identification of non-coding trait-associated SNPs is difficult due to incomplete annotations, imprecise background mutation rates, small effect sizes and a lack of knowledge of their functional roles in complex networks [7].

1.2 The non-coding genome

Coding centric approaches lack the ability to explain complex genotype-phenotype relationships. A broader outlook is needed to attempt to understand these intricate interactions. While there is an increasing consensus on the importance of the non-coding genome in health and disease, making tangible connections is a challenging task.

It was once assumed that all non-coding DNA was “junk” of little-to-no importance. Still, from a perspective of energy conservation, it would not make sense to meticulously conserve all of this “useless information” for no reason. With the development of next generation sequencing (NGS) technologies and bioinformatics, providing faster and more cost-effective analyses, more is being understood on the role of the non-coding genome.

The non-coding genome includes repetitive regions, introns, regulatory elements, and non-coding RNAs (ncRNAs) of varying lengths. ncRNAs are further classified as either long (>200nt) or short (<200nt) depending on the length of their mature products (Figure 2).

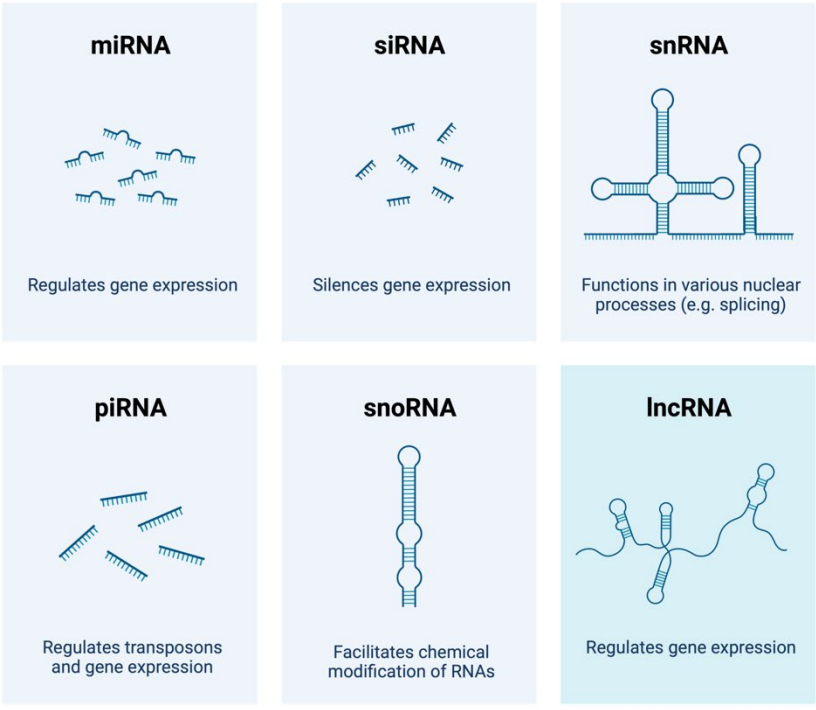


Figure 2. Types of non-coding RNAs. microRNA (miRNA), small interfering RNA (siRNA), small nuclear RNA (snRNA), PIWI-interacting RNA (piRNA), and small nucleolar RNA (snoRNA) are short non-coding RNAs under 200 nucleotides (nt) in length. Long non-coding RNAs (lncRNAs) are over 200nt long (subtypes not shown) [8].

Non-coding regions make up approximately 98% of the human genome and produce 90% of all transcribed sequences. Very little of the observed genetic variation, both within and between species, is found in protein coding genes. GWAS has revealed that the majority of observed diversity is located in transcribed non-coding sequences [9].

The most extensively studied species of ncRNAs today are microRNAs (miRNAs). It has been reported that they regulate over 60% of all protein coding genes, making them the largest class of gene regulators [10]. SNPs in miRNAs can have major impact on their function, and even alter miRNA biogenesis itself through changing the dynamics between the miRNA transcript and the processing enzymes. For example, a G > A substitution in the `mir-30c` gene (rs928508) increases `miR-30c` expression which further modulates chemoresistance in breast cancer [9].

1.2.1 Epigenetics

Gene regulation on a spatiotemporal level is governed by regulatory elements and non-coding RNAs. Neural development, cell differentiation and proliferation, tissue specificity, along with all other dynamic physiological processes, rely on epigenetic regulation. Unsurprisingly, dysregulation of these functions have repeatedly been reported to be associated with cancer [11].

While genetics is the study of the DNA sequence, epigenetics describe the mechanisms that modify gene expression without altering the sequence itself. Such epigenetic modifications can be considered as external forces affecting the rate of transcription.

The three major classes of epigenetic mechanisms are DNA methylation, histone modification, and ncRNA-mediated gene regulation [12]. DNA methylation and histone modifications act at transcription-level, while ncRNAs are post-transcriptional regulators of expression. Moreover, some ncRNAs have the ability to interact with both DNA methyltransferases and histone-modifying complexes, making them both direct and indirect regulators of gene expression [13].

Epigenetic modifications can be mapped with specialised sequencing technologies. Methylation levels can be measured with bisulfite sequencing, and histone modifications by ChIP-Seq (Chromatin immunoprecipitation sequencing) and ATAC-Seq (Assay for Transposase-Accessible Chromatin using sequencing). ChIP-Seq allows for mapping of histone modifications and protein-DNA binding events, while ATAC-Seq provides information on chromatin accessibility.

RNA sequencing (RNA-Seq) enables us to study the transcriptome, both from coding and non-coding regions. The RNAs are isolated and selected for by size. RNA-Seq traditionally refers to the sequencing of long RNAs (mRNAs and lncRNAs) whereas sequencing of short ncRNAs (which utilises the same technology) is customarily called Small RNA-Seq.

1.3 miRNAs

miRNAs are a subclass of short ncRNAs ~22nt in length, first discovered in *C. elegans* in 1993 [14]. They are essential for tissue- and development specific expression due to their versatile yet specific nature allowing rapid fine-tuning of cellular processes [15]. In recent years, many miRNAs have been shown to be major contributors to immunotherapy response and prognosis in cancer [9, 16], and the first clinical application of miRNA targeting was in 2013 [17]. miRNAs can have both tumour suppressing and oncogenic effects, depending on their target genes (Figure 3).

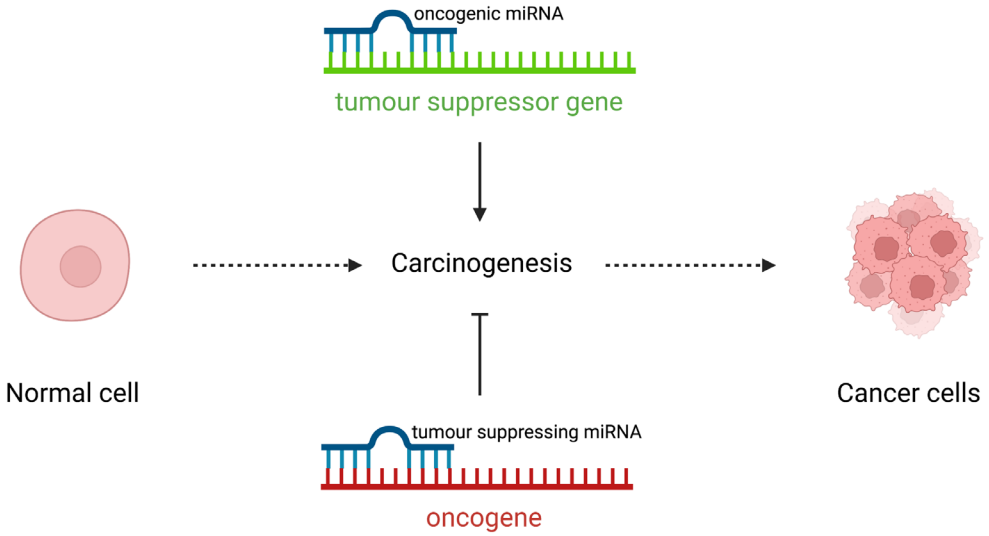


Figure 3. miRNAs can have both tumour suppressing and oncogenic effects. A miRNA can act oncogenic if it inhibits expression of a tumour suppressing gene. Similarly, the miRNA can act as a suppressor by silencing oncogenes [3].

1.3.1 Biogenesis

The canonical pathway is the main maturation pathway of miRNAs (Figure 4A). Most miRNA genes lie in non-coding regions of the genome where they can be several kilobases (kb) in length. Once transcribed, the transcript folds to form a complex hairpin structure known as a primary miRNA (pri-miRNA). While still in the nucleus, the pri-miRNAs are cleaved by the Microprocessor complex (mainly composed of the Drosha enzyme and RNA-binding enzyme DGCR8) to form precursor miRNAs (pre-miRNAs).

The pre-miRNA is transported out of the nucleus by Exportin-5 where it is cleaved by Dicer, releasing the hairpin, and leaving a miRNA/miRNA* duplex. Finally, a mature miRNA combines with one of the Argonaute protein family members AGO1-4 to form the miRNA induced silencing complex (miRISC) [18]. In humans, AGO2 is the most frequently occurring family member. The mechanism(s) behind selecting which miRNA is loaded into the miRISC-complex, and which is degraded, is not fully understood.

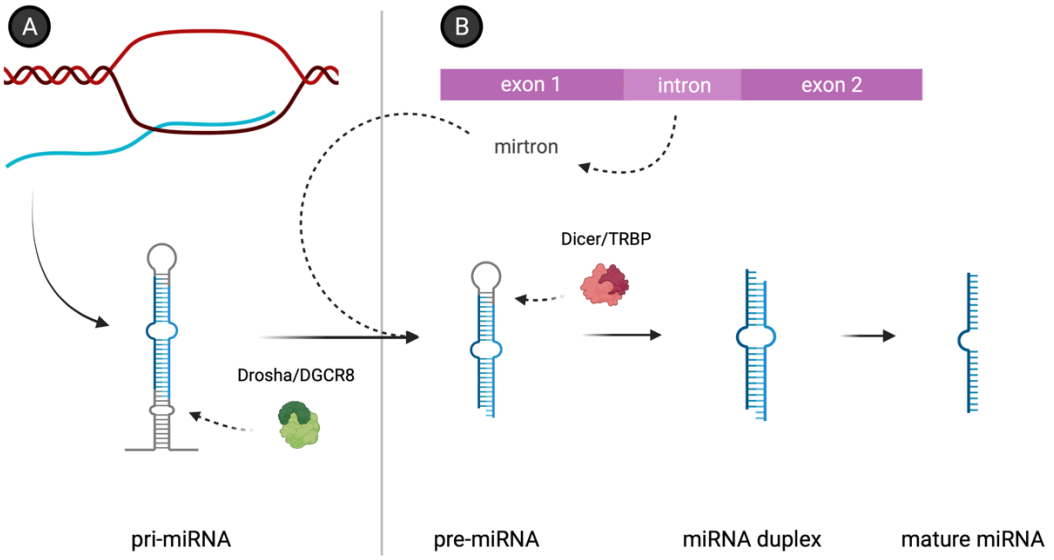


Figure 4. miRNA biogenesis through (a) the canonical pathway, and (b) the non-canonical mirtron pathway [3].

There are also non-canonical pathways where the miRNA bypasses either Drosha processing or Dicer processing. Occasionally, the miRNA gene is found within an intron or in the UTR of a protein coding gene [16]. miRNAs that are located within introns are called *mirtrons*, and bypass Drosha mediated cleavage. Instead, they are by-products of intron slicing and enter the biogenesis pathway as pre-miRNAs (Figure 4B) [19]. miRNAs derived from small nucleolar RNAs (snoRNAs) are matured in a similar fashion.

Very few miRNAs have been reported to bypass Dicer cleavage. One of the few known examples is the `miR-451` transcript where the pre-miRNA is directly loaded into and cleaved by AGO2 [20].

1.3.2 miRNA-mRNA regulatory networks (MMRNs)

miRNAs exert their gene regulatory effect by binding to target mRNAs causing repression of translation through RNA interference.

Canonical binding patterns are formed between the mRNA 3' UTR and the miRNA seed region, generally considered to encompass nucleotides 2-7, counting from the 5' end (Figure 5A). Binding patterns with imperfect seed region complementarity and/or involving other nucleotides in the miRNA sequence are termed non-canonical (Figure 5B). On rare occasions, the miRNA can also bind to the 5' UTR or the coding sequence of its target mRNA [21].

miRNA-mRNA binding mainly results in miRNA-induced gene silencing through translational repression. The exact mechanism remains uncertain, but the consensus is that it involves inhibition of translation by deadenylation, decapping and finally decay of the mRNA [10]. The multistep nature of miRNA-induced silencing makes the process reversible and highly dynamic [22].

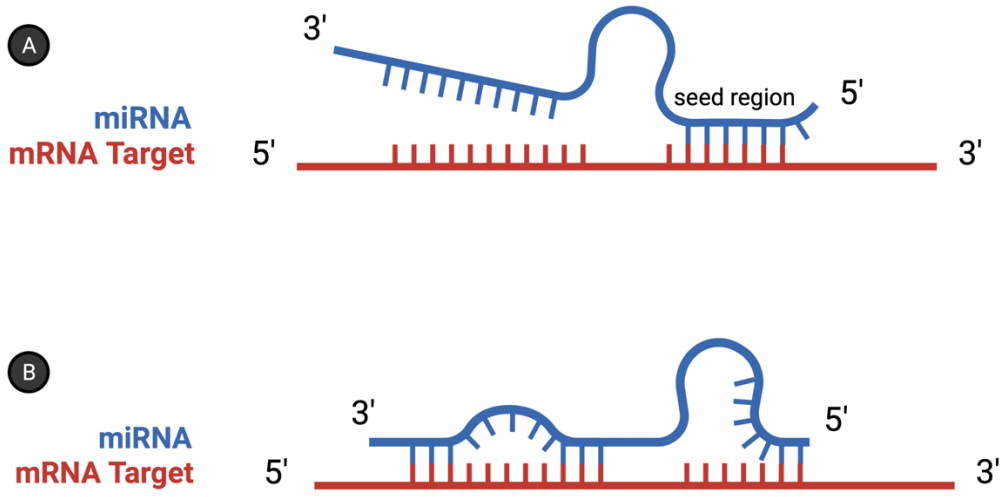


Figure 5. (a) Canonical miRNA-mRNA binding by seed region complementarity. (b) Non-canonical binding pattern [3].

miRNA-mediated cleavage is extensively described in plants, but such behaviour is not well documented in mammals. Human AGO2 does display endolytic activity, such as the cleaving of the `miR-451` transcript, but only under certain conditions. Endonucleolytic cleavage requires near perfect base complementarity between miRNA and mRNA for AGO2 to be situated properly, which is very rare in mammals [23].

Each miRNA can target hundreds of mRNAs, and a single mRNA can be targeted by multiple miRNAs. Seeing there are more than 2600 annotated human miRNAs and ~20,000 coding genes, this allows the creation of highly complex regulatory networks. An MMRN is mainly composed of many weak miRNA-mRNA interactions, underlining the importance of investigating larger populations of miRNAs.

1.3.3 miRBase

There are several miRNA registries available, but the most commonly referenced database is miRBase. MiRBase is an online resource storing miRNA sequences and their annotations, including both the mature miRNAs and the pre-miRNA like hairpin loop. miRBase was established in 2002 and is responsible for miRNA gene nomenclature and the addition of newly discovered miRNAs. 271 species are represented as per the latest release (v22.1), totalling 38,589 pre-miRNA and 48,860 mature miRNA entries – of which 1917 and 2654 are from the human genome, respectively.

With each release, miRBase also reports ‘high confidence’ sets. These are miRNAs that meet certain criteria, indicating their annotations are of high quality. In miRBase v22.1, 26% of human miRNAs are classified as ‘high confidence’. However, there are concerns as far as the lack of coherence between the different versions of these sets. From one year to another, miRNAs have been removed, only to be added back later, contradicting the idea of them being ‘high confidence’.

In addition to the miRNA annotations, miRBase also contains literature references, and information related to both predicted and experimentally validated target genes [24].

1.4 isomiRs

In standard analyses, miRNAs are considered to exist as well-defined features with a precise start and stop position - this is also how they are annotated in miRBase. In reality, they exist as a population of similar but slightly different isoforms, or *isomiRs*, with varying start and stop positions and the possibilities of polymorphisms in the transcribed sequence. isomiRs were first described in 2008 [25].

Coding centric approaches tend to investigate miRNAs in a similar way to genes. While the pri-miRNA transcripts are indeed transcribed by RNA polymerase II like protein coding mRNAs, the miRNAs themselves do not have distinct start and stop codons. Thus, the assumption that they are such clearly defined entities appears to be incorrect, and it has been shown on multiple occasions that isomiRs are of functional importance [26-29]. isomiRs with modifications on the 5' end (having varying start positions) are considered especially important due to alteration of the seed region, as they have the potential to cause major changes in mRNA targeting. Nevertheless, they are neglected in most miRNA studies.

1.4.1 mirGFF3

As the amount of research on isomiRs increased, there became an abundance of different reporting standards, making it difficult to compare results across studies. To address this, the mirGFF3 format was introduced in 2019 [30]. For this project it was considered prudent to conform to this standard for the output to be reproducible and reusable in the future.

mirGFF3 is an adaptation of the already well established GFF3 format used for describing genes (full definition available at https://github.com/miRTop/mirGFF3). It is a tab delimited file with 9 columns, where the first 8 columns contain the pre-miRNA name, source (database), feature type, start/end position, score (optional), strand and phase (optional):

```
MI0000077:hsu-mir-21  miRBasev22.1u  isomiR  59841266 59841337  .  +  .
```

Finally, column 9 contains a list of additional feature attributes in the format tag=value, and is used for storing isomiR specific details:

```
Read=TAGCTTATCAGACTGATGTTG;UID=iso-21-VIV6OYIN0;Name=hsu-miR-21-
5p;Parent=MI0000077;Variant=iso_3p:-1;Cigar=21M;Expression=2564
```

Structure of the attribute string and framework of the mirGFF3 variant classification are described in detail in the Methods (section 2.2.1). Here, we provide a simple introduction to the general layout of the format.

When describing characteristics of isomiRs, some terms also need to be established. Foremost, when speaking of a *miRNA,* we are referring to the annotated genomic feature (for example in miRBase). A *variant* refers to one of a set of categorised modifications to a miRNA, such as the deletion of a 3' nucleotide. An *isomiR* is a specific miRNA variant. For instance:

miRNA: `miR-21-5p`

Variant: `iso_3p:-1`

isomiR: `miR-21-5p;iso_3p:-1`

A group of isomiRs corresponding to a certain miRNA is referred to as an isomiR population, and an isomiR profile refers to the characteristics of the distribution of this population.

Finally, *trimming variants* are isomiRs that are simply shorter or longer than the reference miRNA, implying that the changes are due to imprecise trimming of the pre-miRNA hairpin (Figure 4). Trimming variants can, for instance, be caused by variation in Dicer cleavage, or by other post-transcriptional mechanisms.

1.5 Challenging the coding centric view

Despite the increasing knowledge on the of role the non-coding genome and gene regulation, the focus of medical research remains coding centric.

The dilemma can be illustrated by a classic trait widely known to be heritable - human height (Figure 6). Heritability in this context is a measurement of the proportion of phenotypic variance that can be attributed to genetic factors [31]. Human height has an estimated heritability of approximately 80%, however, less than 4% can be explained by significant GWAS-hits. The number increases to 45% when considering all SNPs regardless of significance, which still leaves a large amount of unexplained variance [32].

This phenomenon is called missing heritability and describes the discrepancy between observed heritability and the amount that can be explained by genotype data. Such observations have further challenged the view on polygenic traits, suggesting that the genetic architecture of complex traits is truly omnigenic – meaning all expressed genes have the ability to affect each other [33]. Shifting focus to the non-coding genome and epigenetic regulators could help to further clarify the association between genotype and phenotype.

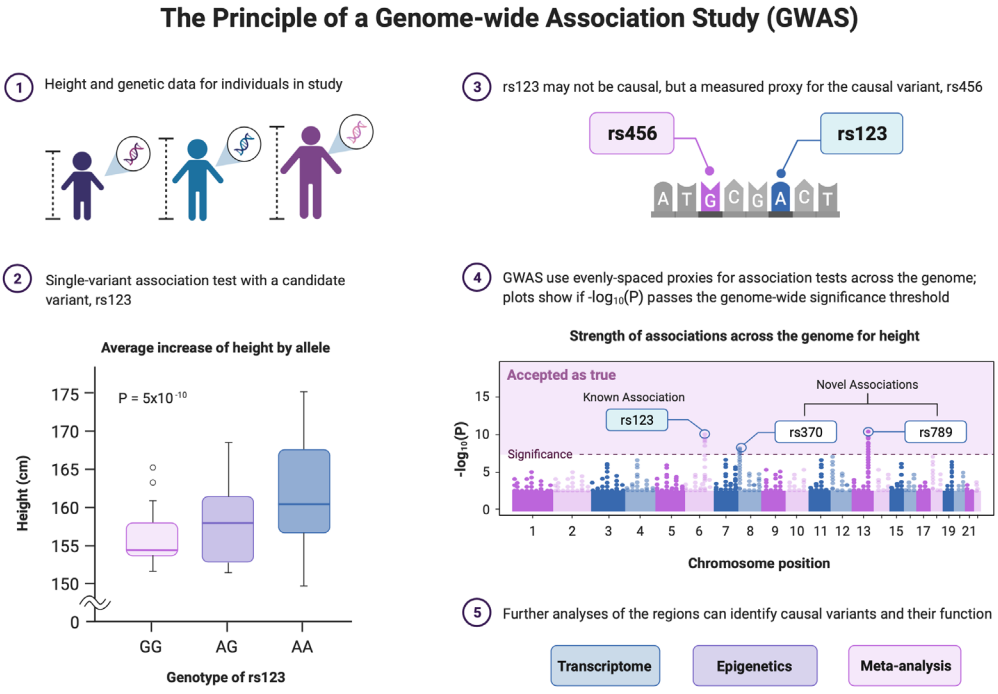


Figure 6. The principle of a genome-wide association study (GWAS). Height measurements and sequencing data are obtained from the individuals, and correlation is measured. P-values are used to determine the significance of a given variant [34].

1.6 Motivations and aims

Evidence suggests that miRNA signatures could have clinical applications as biomarkers for diagnostic and prognostic purposes, and their potential for regulating whole pathways make them attractive therapeutical targets [21].

It has been shown that cancers can be distinguished based on their isomiR profile alone [35]. However, while these studies demonstrate that different cancers display different isomiR profiles, they do not identify what (i.e., which isomiRs) is driving the classification.

In this work, we perform a comprehensive investigation of how isomiR populations vary among different conditions. Specifically, the project will investigate how isomiR populations vary between healthy and cancer patients.

The primary aim of the study is to

1) Develop software to identify, classify, and report isomiRs

And secondly

2) Visualise isomiR distributions

to provide a foundation for the characterisation of isomiR populations in disease and aid in a better understanding of the genetic architecture of complex conditions.

2 Methods

A wide selection of tools has been utilised throughout the project. The data analysis pipeline is a combination of software developed as part of this work and pre-made software. Post-processing and visualisation methods were developed independently as part of this work. All software and version specifications are available on GitHub at https://github.com/CBGOUS/BagEnd. Subsequent mentions of GitHub refer to this repository.

2.1 FAIR data

With the emerging wealth of data being produced in academia and industry, there is a pressing need for underlying infrastructure and data management guidelines. To address this, the FAIR guiding principles were introduced by Wilkinson and colleagues in 2016 [36]. They are:

1) **F**indability
2) **A**ccessibility
3) **I**nteroperability[1]
4) **R**eusability

As described by Wilkinson, the quality and impact of a publication should be "[…] a function of its ability to be accurately and appropriately found, re-used, and cited over time […]" (p.3).

FAIR data is usually spoken of in the context of research output data and data management. However, these principles also apply to the tools, workflows and pipelines needed to produce the output data. Without full transparency in all components of the research process, the output data cannot entirely be in accordance with the FAIR-principles.

The choice of output format, code structure and post-processing tools have been made with these considerations in mind. Code readability and lexicon quality (comments and identifier names like methods and variables) have been prioritised over peak performance and minimalism. Trying to understand, debug and use overly compacted code, however effective, can end up being more time consuming than its slower counterpart. This is well documented to be a large cost of any software [37].

[1] Interoperability – the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort [30].

2.2 NGS pipeline

The NGS analysis pipeline used in this work was developed at the Computational Biology Group at Oslo University Hospital (CBGOUS) and consists of more than 30 steps made to process Small RNA-Seq data. The applications of the pipeline exceed far beyond the scope of this project, therefore only the steps relevant to this analysis will be described.

The pipeline is written in Java and runs as an executable jar file. Running the program requires three input files, which are specified in Table 1, in addition to a log configuration file. In brief, the Data file contains the file names, the Pipeline file specifies which analysis steps to perform on the data files, and the Configuration file supplies step specific parameters and paths to the necessary tools. Full examples of each file type can be found on GitHub.

Table 1. Input files to the NGS analysis pipeline.

Argument	File type	File format	Description
-c	Configuration file	`yaml`	Paths to the reference data and software needed to perform the steps. Step specific parameters, such as Bowtie alignment options, are also specified here.
-d	Data file	`tsv`	Tab separated file with names of fastq-files and necessary metadata such as grouping information/condition.
-p	Pipeline file	`json`	File containing the analysis steps to perform, input/output folder names, and a parameter string. The parameter string pulls the step specific parameters from the Configuration file.

The data processing of this project is performed across 8 steps and is shown in Figure 7. Steps 1 through 5 carry out the mapping of the sequences to the reference genome along with the necessary pre-processing. The input files are unzipped (step 1) before the adapter sequences are trimmed from the raw reads (step 2). The trimmed reads are further converted from fastq to fasta format, and all duplicate reads are collapsed to unique sequences, adding up and storing the count information in the fasta header-line (step 3).

The unique reads are further filtered for known contaminants and ribosomal RNA sequences (step 4), and finally the remaining sequences are mapped to the reference genome. Step 4 and 5 both use Bowtie [38] for mapping, with a mismatch acceptance of 1 and 0, respectively.

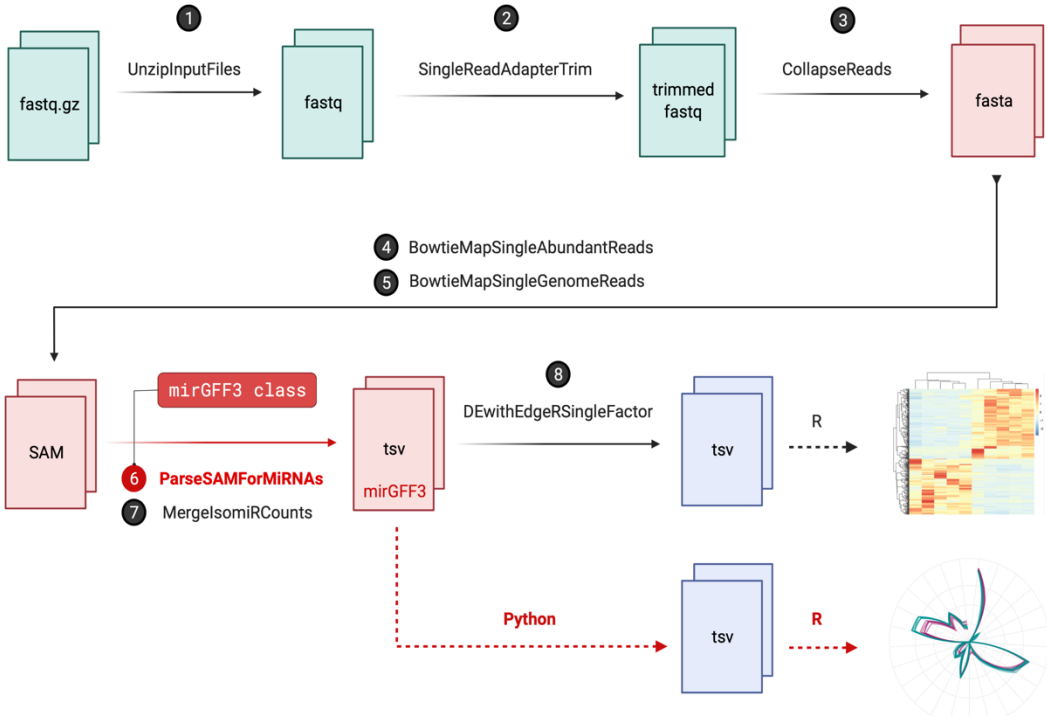


Figure 7. Flowchart of the steps in the NGS analysis pipeline. Not shown is the step for performing quality control checks with FastQC [39], which can be implicated before trimming to aid in identification of the adapter sequence, and/or after trimming to validate the quality of the data. Dashed lines indicate work done outside of the main NGS pipeline. Contributions from this project are highlighted in red [40].

The isomiR classification is performed in step 6. Step 6 processes the mapped miRNA reads to determine isomiR content for each parsed file (sample). The details of the classification process are described further in the next section. After classification, the isomiR count information for each sample is run through step 7 and joined into a single merged count file. The output files from steps 6 and 7 are shown in Table 2.

Different users of the software will likely have different needs, and the variety of output files and formats aims to meet as many as possible. Sample specific files are marked with sample ID (filename), and project specific files are marked with project ID (in this context, the project equals all samples in the input Data file).

The 8th and final step of the NGS analysis pipeline constructs an R-script to perform differential expression analysis on the merged count file using edgeR from Bioconductor [41].

For this work, the output files `mirGFF3` and `merged_isomir_counts` are used for isomiR visualisation and differential expression analysis, respectively.

Table 2. Output files from the NGS analysis pipeline. m = number of isomiRs, n = number of samples.

Step	File name	File format	Description
6	SAMPLEID	`mirGFF3`	All detected isomiRs are written out in the mirGFF3 format.
6	SAMPLEID.trim.clp.gen.isomircounts	`tsv`	Tab separated file with *m* rows and 2 columns. First column is the isomiR name, second column is the counts of this isomiR in the current sample.
6	SAMPLEID.iso_pretty	`tsv`	Tab separated file visually representing the isomiRs and their dispersion.
6	PROJECTID_uniq_isomirs	`fasta`	All unique isomiRs detected across all samples written in fasta format.
6	PROJECTID_uniq_isomirs	`tsv`	Tab separated file with *m* rows and 2 columns containing all isomiRs detected across all samples. First column is the isomiR name, second column is the corresponding sequence.
7	PROJECTID.merged_isomir_counts	`tsv`	Tab separated file with *m* rows and *n+1* columns. First column is the isomiR name, and subsequently one column for each sample.
7	PROJECTID.trim.clp.gen.isomircounts_wseq	`tsv`	Same as above, including an extra column with the sequence for each isomiR.

As an example, the layout of the `pretty plot` is shown below. Rows one and two show the miRNA annotation information and total count number, respectively. The subsequent rows contain the sequence, MD-string, count number, and abundances for each variant of this miRNA (as percentage of the total population).

This ‘plot’ allows for a simple, visual inspection of the results in plain text format without the need for post-processing.

```
hsu-miR-21-5p|MIMAT0000076 : chrMI0000077:hsu-mir-21 + 8 --> 29 (+) : UAGCUUAUCAGACUGAUGUUGA
Total Counts = 36682

       hsu-miR-21-5p      TAGCTTATCAGACTGATGTTGAC        MD:Z:23      15737     42.9%
       hsu-miR-21-5p      TAGCTTATCAGACTGATGTTGA         MD:Z:22      10919     29.77%
       hsu-miR-21-5p      TAGCTTATCAGACTGATGTTGACT       MD:Z:24       6808     18.56%
       hsu-miR-21-5p      TAGCTTATCAGACTGATGTTG          MD:Z:21       2564      6.99%
       hsu-miR-21-5p      TAGCTTATCAGACTGATGTT           MD:Z:20        564      1.54%

                                        .
                                        .
                                        .
```

2.2.1 The mirGFF3 class

The classification of isomiRs is performed in the `mirGFF3` class, which has been developed as a part of step 6 of the NGS pipeline (Figure 7).

Calling a step in the `json` file requires four pieces of information; the step name, input folder, output folder, and a parameter string. For step 6, the `json` entry is:

```
        {"step":"ParseSAMForMiRNAs"," inputFolder":"inputfolder",
"outputFolder":"outputfolder",   "parameterString":"parseIsoUnfiltered"}
```

The parameter string is used to pull a specific set of parameters from the `yaml` file, which are used to customise the functionality of the pipeline. The `yaml` entry for step 6 containing the parameters used in this project is the following:

```
ParseSAMForMiRNAs:
    Required:

    Optional:
        parseIsoUnfiltered:   " --bleed=2, --mirbase_version=22.1u,
        --ref_genome=hsu, --min_counts_per_million=0,
        --analyze_isomirs=true, --group_by_seed_region=false,
        --seed_string=2-7"
```

For the `parseIsoUnfiltered` parameter string, the required flags are `bleed`, `mirbase_version`, and `ref_genome`. The reference genome "`hsu`" used in this work will be explained further in the next section.

Additionally, the user can employ a filtering option (`min_counts_per_million`) to discard low counts and specify whether the software is to analyse isomiR content or not (`analyze_isomirs`). If the `group_by_seed_region` flag is set to `true`, isomiRs will be grouped according to the seed region (specified through `seed_string`). The value of this is questionable, but the argument can be made that isomiRs with identical seed regions may have similar targets.

The `bleed` value states how many positions the isomiR of interest is allowed to diverge from the reference miRNA's start- and stop positions. For this work a threshold of 2 has been used, and all reads starting or ending more than two positions from the mature reference miRNA are discarded.

As an example, we consider the following reads mapped to the miRNA `miR-21-5p`:

```
a)      TGTCGGGTAGCTTATCAGACTGATGTTGACTGTT
b)            TAGCTTATCAGACTGATGTTGA
c)            TAGCTTATCAGACTGATGTTGAC
d)              GCTTATCAGACTGATGTTGA
e)          GGGTAGCTTATCAGACTGATGTTGA
f)            TAGCTTATCAGACTGATGT
```

Here, read **a)** shows a portion of the `miR-21-5p` hairpin precursor, with the annotated miRNA in red letters and the bleed region highlighted in yellow. Read **b), c)** and **d)** have start- and stop positions within the threshold and will be retained. Read **e)** and **f)** are discarded as they fail to meet the bleed cut-off due to a 3-nucleotide extension and 3-nucleotide deletion, respectively.

The full range of possible 5'/3' isomiR variants is shown in Figure 8. The modifications include deletions, template extensions, non-template extensions, and the compound event of a deletion followed by a non-template extension. Template extensions refer to extensions of the mature miRNA sequence that match the hairpin precursor nucleotide in the same position (for instance, read **c)** and **e)** from above). For this project, only the trimming variants will be considered (green and red events).

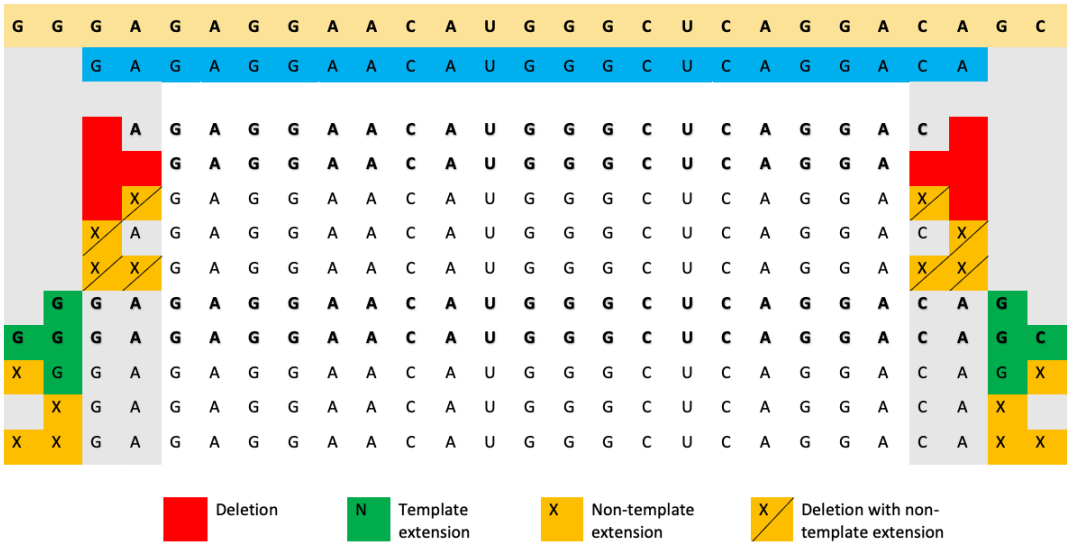


Figure 8. All possible isomiR trimming variants considering a bleed of 2. The first row shows a portion of the pre-miRNA hairpin sequence, and below it the reference miRNA is shown in blue. Trimming variants are the variants with only deletions and/or template extensions (red/green events in bold).

IsomiR trimming variants in the mirGFF3 format are in the form `iso_5p:+/-N` and `iso_3p:+/-N`. A full description of variant nomenclature can be found in Table 3.

For conciseness, from here on out they will be referred to as `5p:+/-N` and `3p:+/-N`, respectively. `N` indicates the number of positions the end- or start position is shifted to the right. Revisiting `miR-21-5p` reads **a)** through **d)** from the previous page, they correspond to the following variants:

```
a)      TGTCGGGTAGCTTATCAGACTGATGTTGACTGTT
b)        TAGCTTATCAGACTGATGTTGA              ref
c)        TAGCTTATCAGACTGATGTTGAC             3p:+1
d)          GCTTATCAGACTGATGTTGA               5p:+2
```

The end position of read **c)** is shifted one position to the right due to a 3' extension, giving it the variant name `3p:+1`. Read **d)** has 2 deletions on the 5' end, i.e., the start position is shifted 2 positions to the right, which equals variant `5p:+2`.

In addition to a variant name, all isomiRs receive their own unique identifier (UID) of the form:

`iso-NN-C{N}`

where `NN` corresponds to the length of the isomiR, and `C{N}` to the encoded nucleotide sequence. The UID is encoded from the sequence according to the MINTbase framework [42]. Each 5-, 4-, 3-, 2- and 1-mer of nucleotides can be represented by one or two letters that are not A, G, T or C. As an example, the reference variant of `miR-21-5p` translates to:

`TAGCTTATCAGACTGATGTTGA  ->  VIV6OYINL`

And the UID-string for the `miR-21-5p` reference miRNA becomes:

`iso-22-VIV6OYINL`

Finally, the information is combined into the full mirGFF3 attribute string:

```
Read=TAGCTTATCAGACTGATGTTGA;UID=iso-22-VIV6OYINL;Name=hsu-miR-21-
5p;Parent=MI0000077;Variant=ref;Cigar=22M;Expression=10919
```

The attribute string is a semicolon-delimited string composed of the read sequence, UID-string, miRNA name, miRBase accession number for the parent pre-miRNA, isomiR variant, cigar string, and expression (count) number. The cigar string originates from the SAM file and specifies the matches (or mismatches) between the read and the reference. For example, “22M” stands for “22 matches” between the read and the reference. Cigar strings also contain information on mismatches, deletions, insertions etc., but for this project no mismatches have been allowed in the alignment.

Table 3. isomiR variants in the mirGFF3 format.

Variant	Description
`iso_5p:+/-N`	+ indicates the start is shifted to the right, - indicates the start is shifted to the left. N is the number of nucleotides it differs.
`iso_3p:+/-N`	+ indicates the start is shifted to the right, - indicates the start is shifted to the left. N is the number of nucleotides it differs.
`iso_5p:+/-N,iso_3p:+/-N`	In case of modifications on both 3’ and 5’ ends, variants are separated with “,”
`iso_add3p:N`	Number of non-template nucleotides added at 3’ end.
`iso_add5p:N`	Number of non-template nucleotides added at 5’ end.
`iso_snv_seed`	SNV in nucleotides 2-7
`iso_snv_central_offset`	SNV in nucleotide 8
`iso_snv_central`	SNV in nucleotides 9-12
`iso_snv_central_supp`	SNV in nucleotides 13-17
`iso_snv`	SNV in any other nucleotide

2.3 The reference genome

Species in miRBase are represented by a three- or four-letter code such as `hsa` for *homo sapiens* and `mmu` for *mus musculus*. This species ID needs to be specified in the Configuration file (Table 1) under *host name*. The host name is used to locate the annotation- and sequence files for the reference genome:

1) Genome annotation file (`gff3`)
2) pre-miRNA hairpin sequences (`fasta`)
3) Mature miRNA sequences (`fasta`)

A custom subset of miRBase has been used for this project, being deemed most fit for the purpose. The aim of this project is to study isomiR populations and their dispersion in human tissue rather than determining their origin. Since many miRNAs are present in multiple copy numbers throughout the genome, these have been reduced to unique sequences only. For convenience and speed, the start- and end position for mature miRNAs are given relative to their pre-miRNA hairpin and not as absolute genome coordinates.

When working with custom subsets of species, a new host name should be given to avoid confusing it with the original. For this project, the host name is set to `hsu` for *homo sapiens unique sequences*. The reference genome and sequence files should then be named accordingly, as shown in Table 4. The miRNA names must also be changed to match the new host name:

```
hsa-miR-21-5p -> hsu-miR-21-5p
```

This is due to the way the pipeline indexes and maps sequences based on the input host name. Before running the pipeline, the genome index for the specific host must be built with Bowtie [38].

Table 4. Annotation- and sequence files for the reference genome.

File	File name	Comment
Genome reference	hsu_genome.gff3	Adapted from miRBase v22.1: `hsa.gff3`
pre-miRNA sequences	hsu_hairpin.fa	Adapted from miRBase v22.1: `hsa_hairpin.fa`
Mature miRNA sequences	hsu_mature.fa	Adapted from miRBase v22.1: `hsa_mature.fa`
Bowtie index	hsu_genome.[a-z].[0-9].ebwt	`> bowtie-build -f -r hsu_hairpin.fa hsu_genome`

2.4 Data collection

Raw reads in fastq format were downloaded from the Sequence Read Archive (SRA) at https://www.ncbi.nlm.nih.gov/sra with the SRA toolkit command `fastq-dump` [43]. Inclusion criteria for the datasets were:

1) Raw Small RNA-Seq data
2) Patient-derived samples (not cell lines)
3) Patients had not undergone treatment
4) Minimum of 5 disease samples + 5 controls

3 SRA projects (SRP) containing samples from 4 different conditions meeting these criteria were selected. The conditions were glioblastoma (GBM), lung adenocarcinoma (LAC), colorectal cancer (CRC) and prostate cancer (PC). The raw `fastq` files were inspected to determine the adapter sequences. `FastQC` was then performed on all samples after adapter trimming to ensure successful trimming and satisfactory read quality. Two of the conditions' samples were sourced from plasma microvesicles, and two from tissue samples. Further details are provided in Table 5.

Table 5. SRA datasets used in the project.

Condition	Source	SRP	3' adapter sequence
CRC (n=200)	Plasma microvesicles	SRP061240	5'-AGATCGGAAGAGCACACGTCT-3'
PC (n=72)	Plasma microvesicles	SRP061240	5'-AGATCGGAAGAGCACACGTCT-3'
GBM (n=5)	Tumour/frontal cortex	SRP063390	5'-TGGAATTCTCGGGTGCCAAGG-3'
LAC (n=10)	Tumour/adjacent normal tissue	SRP359604	5'-TGGAATTCTCGGGTGCCAAGG-3'

2.5 Post-processing and visual representations

2.5.1 Data wrangling

The `mirGFF3` output files from step 6 in the NGS pipeline were further processed in Python to prepare the data for visualisation. Expression values were normalised by counts per million (CPM) and log2-transformed. A summary file was created for each dataset, where counts for each isomiR variant in a sample was summed up across all miRNAs. The fraction occupied by each variant was then computed.

Based on the differential expression report from step 8, miRNAs of interest were extracted from the datasets, and their expression values CPM normalised and log2-transformed. The fraction occupied by each variant was then computed.

2.5.2 Visualisation

Heatmaps of the log2fold-changes from differential expression analysis were generated in R using `pheatmap` [44].

Additionally, radar plots of isomiR expression and distributions were created in R using `ggplot2` [45]. The foundation of the plot is a line plot with isomiR variants on the `x`-axis and counts on the `y`-axis. The plots are then transformed from cartesian (`x,y`) to polar (r,θ) coordinates where `r` is the counts, and the variants are the θ values.

There are 25 variants in total, including the reference. In the radar plots, variants are arranged according to the unit circle principle, although with some necessary modifications (

Figure 9). Starting with the reference variant at the 90° angle, the rest of the variants are situated as follows:

4th quadrant (-,+): - Single end 3’ extensions - 5’ deletion + 3’ extension	1st quadrant (+,+): - Single end 5’ extensions - 5’ extension + 3’ extension
3rd quadrant (-,-): - Single end 5’ deletions - 5’ deletion + 3’ deletion	2nd quadrant (+,-): - Single end 3’ deletions - 5’ extension + 3’ deletion

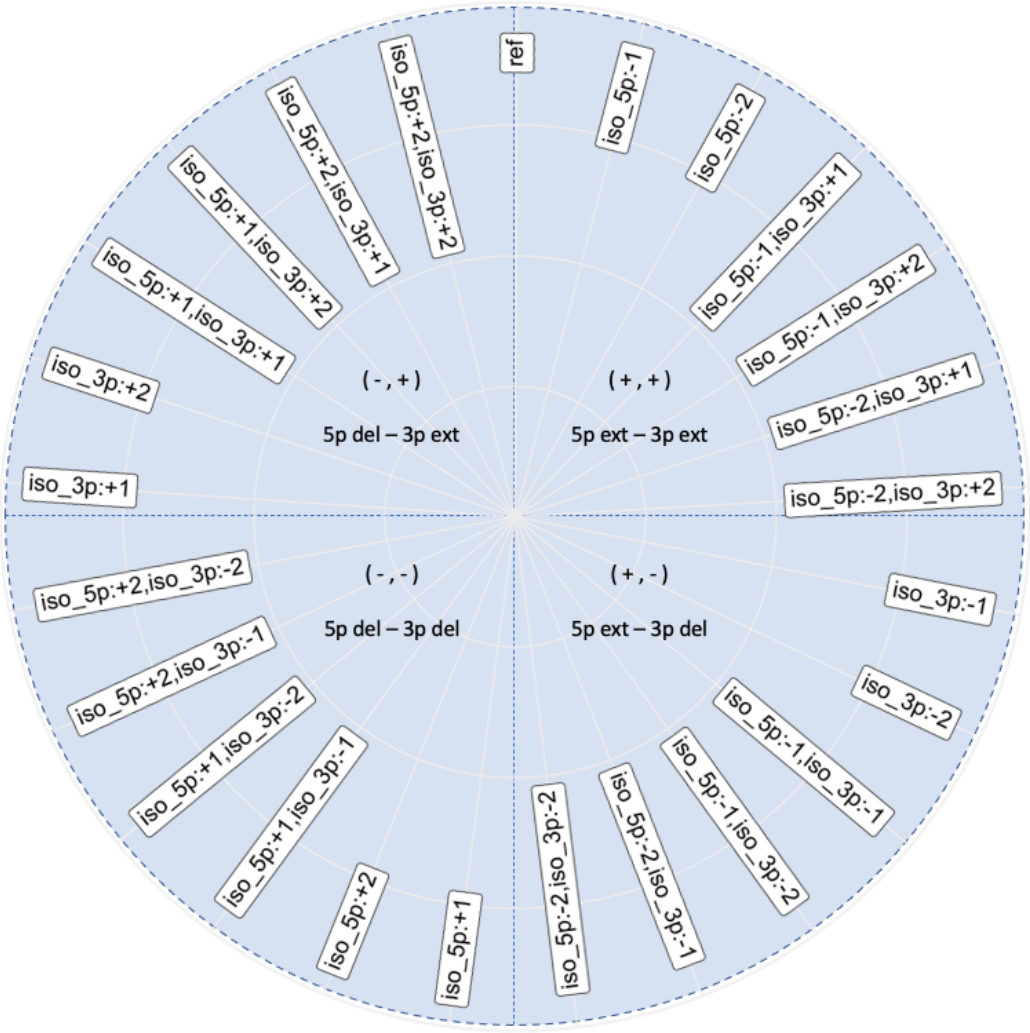


Figure 9. Structure of the isomiR radar plots. The variants are arranged according to the unit circle principle where quadrants 1-4 are (+,+), (+,-), (-,+) and (-,-). + and – correspond to extensions and deletions, respectively. Due to there being 25 variants, the plots could not be perfectly divided into quadrants.

Two different versions of these plots will be presented: expression plots and fraction plots. The expression plots have log2-transformed normalised counts as `r` values and show the volume of the isomiR population and how it is distributed among the different variants. The fraction-plots show the fraction, or proportion, that each of the variants represent in the given population.

2.6 Target prediction using miRAW

Target predictions were performed for a selection of isomiRs using `miRAW` [46]. `miRAW` is a deep learning-based method for predicting miRNA targets that is trained by analysing the whole transcript of the miRNA and the target gene 3' UTR.

23 isomiRs were compared to ~19,700 3' UTRs, and `miRAW` predicts probable binding regions and calculates the free binding energy for each isormiR/target gene pair. The prediction results were filtered by the prediction probability (==1) and predicted binding free energy (< -25 kcal/mol). The network was visualised with Cytoscape [47].

3 Results

3.1 Software

The resulting software is a Java program that can successfully identify and report a wide array of isomiRs, complete with Python and R scripts to post-process and visualise the results.

Step 6 of the NGS Pipeline has been significantly rewritten and modifications were made to the reporting output. Primarily, the Java class mirGFF3String has been created to carry out the identification and classification of isomiRs in NGS data. One of the key methods in this class checks for SNVs and classifies them according to their position in the isomiR sequence (Table 3). Additional methods perform the checks for 5' deletions, 5' extensions, 3' deletions, and 3' extensions, and combine them if multiple events have occurred. Having these methods separated makes it possible to accommodate for all 5'/3' events and combinations thereof, as well as making the code more readable for subsequent updates. Finally, one method performs the translation of the sequence in question into its UID-string.

Although only trimming variants have been analysed in this work, the program is fully able to handle all variants described in the mirGFF3 format. It has been tested for all 5'/3' variant combinations shown in Figure 8, in addition to SNV variants. The program successfully classifies them according to the naming convention in Table 3.

3.2 Differential expression

Heatmaps of log2fold-changes between conditions and their respective controls are presented in Figure 10 and Figure 11. The columns (miRNAs) and rows (variants) are sorted by the frequency of which they occur in the differential expression analysis. isomiRs not differentially expressed are coloured grey.

Overall, the variants most susceptible to differential expression in all conditions were `ref`, `3p:-1`, and `3p:-2`. As can be seen in Figure 10 and Figure 11, there are distinct bands for variants `ref`, `3p:-1` and `3p:-2,` which all cluster at the top. There is also a visible but slightly less prominent band for `3p:+1`.

3' modifications are markedly more common than 5' modifications, and 3' deletions more common than 3' extensions. The only two variants never seen differentially expressed were `5p:-2,3p:+1` and `5p:-2,3p:+2`. These are variants with simultaneous extension events on both ends making them 3 and 4 nucleotides longer than the reference, respectively.

There seems to be no preference for any particular variant to be exclusively up- or downregulated in the disease state. If a variant (such as `3p:-1`) is repeatedly differentially expressed across many miRNAs, the expression change is typically seen in both directions. However, the overall change in an isomiR population for a specific miRNA tends to be in the same direction, i.e., the isomiRs are all increasing, or decreasing.

In rare cases, some variants of a miRNA are seen to deviate from this trend. For instance, `miR-24-5p` and `miR-219a-5p` each display both up- and downregulated isomiRs in GBM. The former shows a log2fold-change of *-1.5* for its `3p:-1` variant and *+5.2* for `5p:-2,3p:-2` (2 nucleotide extension on 5' and 3' end). This is also one of the rare cases where `5p:-2,3p:-2` is differentially expressed, which only occurred 7 times across all datasets and miRNAs.

Of the differentially expressed isomiRs, 332 were unique to GBM, 128 to LAC, 10 to CRC and 28 to PC samples.

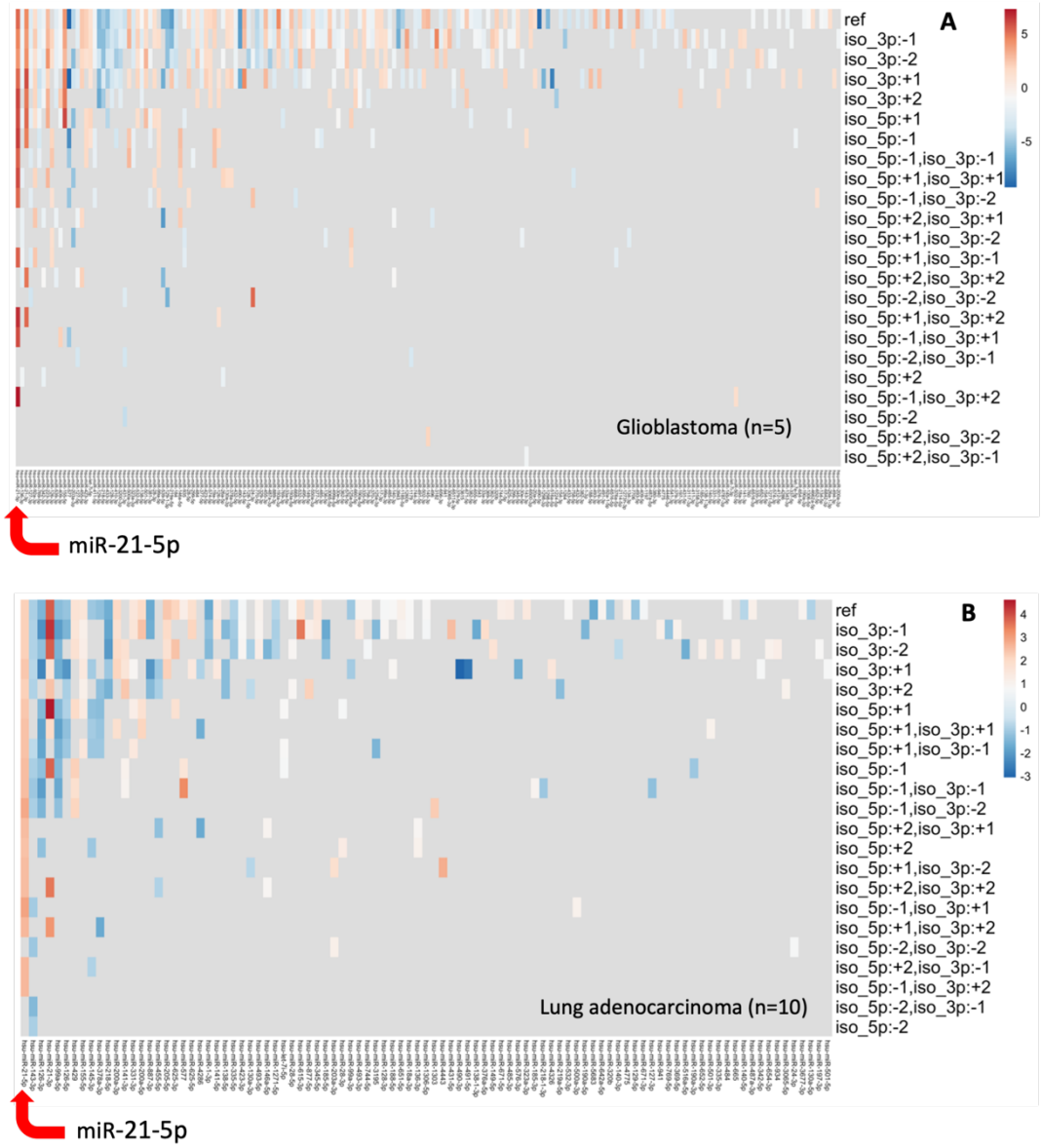
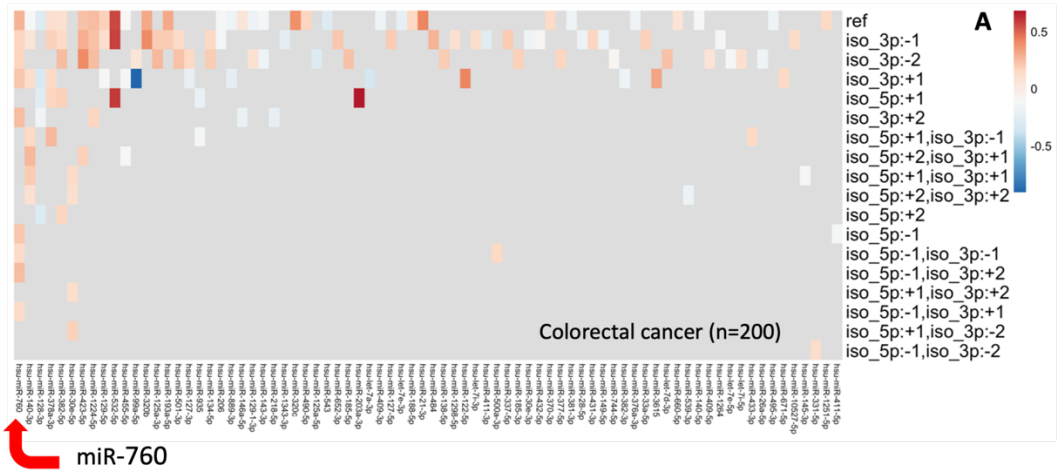


Figure 10. Log2fold-changes for (a) GBM and (b) LAC. The columns (miRNAs) and rows (variants) are sorted by the frequency of which they occur in the differential expression analysis. Red arrows mark `miR-21-5p`, the most frequently differentially expressed miRNA in both GBM and LAC.



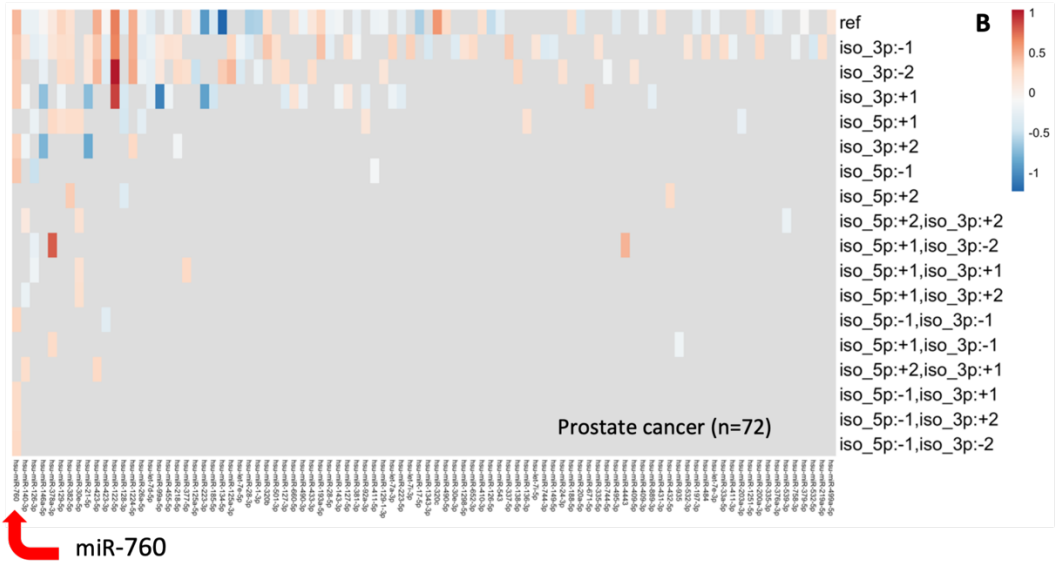


Figure 11. Log2fold-changes for (a) CRC and (b) PC. The columns (miRNAs) and rows (variants) are sorted by the frequency of which they occur in the differential expression analysis. Red arrows mark `miR-760`, the most frequently differentially expressed miRNA in both CRC and PC.

3.3 isomiR profile patterns

The secondary aim of the project was to visualise the identified isomiR populations. A few examples of the applications of the visualisation will be discussed.

Figures 12-15 show the volume and distribution of the 25 isomiR variants across all miRNAs in different conditions. The expression plots (12A-15A) show that expression levels of miRNAs and their variants vary with both tissue type and condition. Predominantly, lung tissue and LAC (Figure 13A) samples display very high miRNA expression values, up to 4 times greater than GBM (Figure 12A) and 1600 times greater than CRC (Figure 14A) and PC (Figure 15A) samples. The population of isomiRs also seems to be more diverse in brain- and lung tissue.

Further, it shows that the most commonly occurring isomiRs are single nucleotide 3' trimming variants (`3p:-1` and `3p:+1`). Variants with 5' modifications are, comparably, rarely seen. They make up a small share of the population (on average, all 5' variants combined accounted for ~2.6% of the population) and are seldom differentially expressed.

In GBM samples and the respective controls the three top variants (`ref`, `3p:-1`, and `3p:+1`) collectively accounted for roughly the same share of the population (87% vs 87.2%, respectively). However, the individual proportions were different, being 59%, 18%, and 10% of the population for GBM versus 75%, 16%, and 2.2% for controls. For LAC samples, the changes in composition were less noticeable, with a 3%, 8%, and 4.3% difference for `ref`, `3p:-1` and `3p:+1`, respectively. The changes in composition between CRC samples, PC samples and controls were negligible (Table 6).

Table 6. Average % of the population accounted for by the main variants (all miRNAs).

	`ref`	`3p:-1`	`3p:+1`	**total**
GBM/control	59%/75%	18%/10%	10%/2.2%	87%/87.2%
LAC/control	57%/54%	28%/36%	7.9%/3.6%	92.9%/93.6%
CRC/control	45%/46%	39%/38%	2.4%/2.6%	86.4%/86.6%
PC/control	45%/46%	40%/38%	2.1%/2.6%	87.1%/86.6%

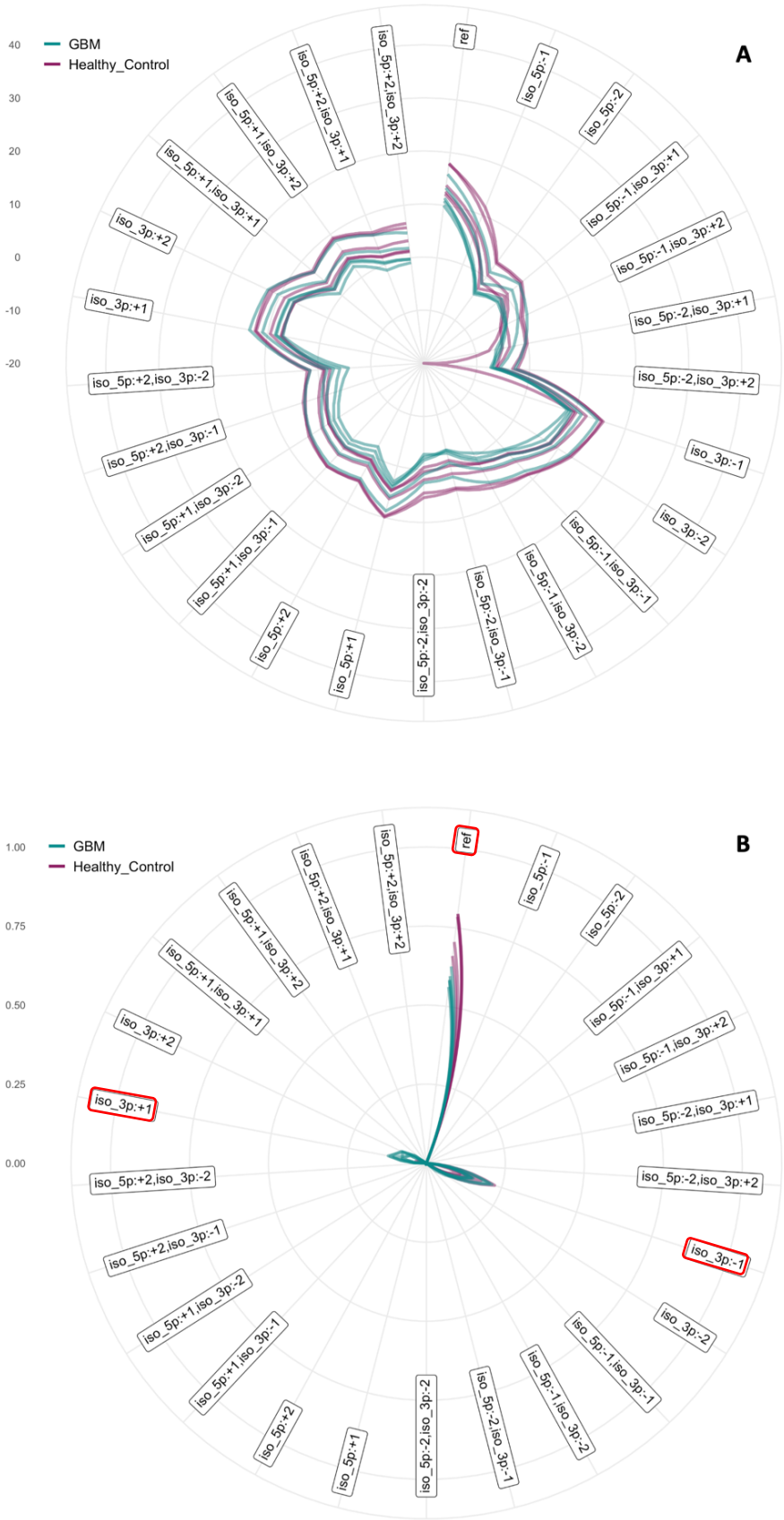


Figure 12. (a) Log2-expression and (b) composition of the isomiR populations across all miRNAs in GBM samples. The reference and single nucleotide 3' extension/deletion variants (in red) make up 87% and 87.2% of the isomiRs in disease state versus control, respectively.

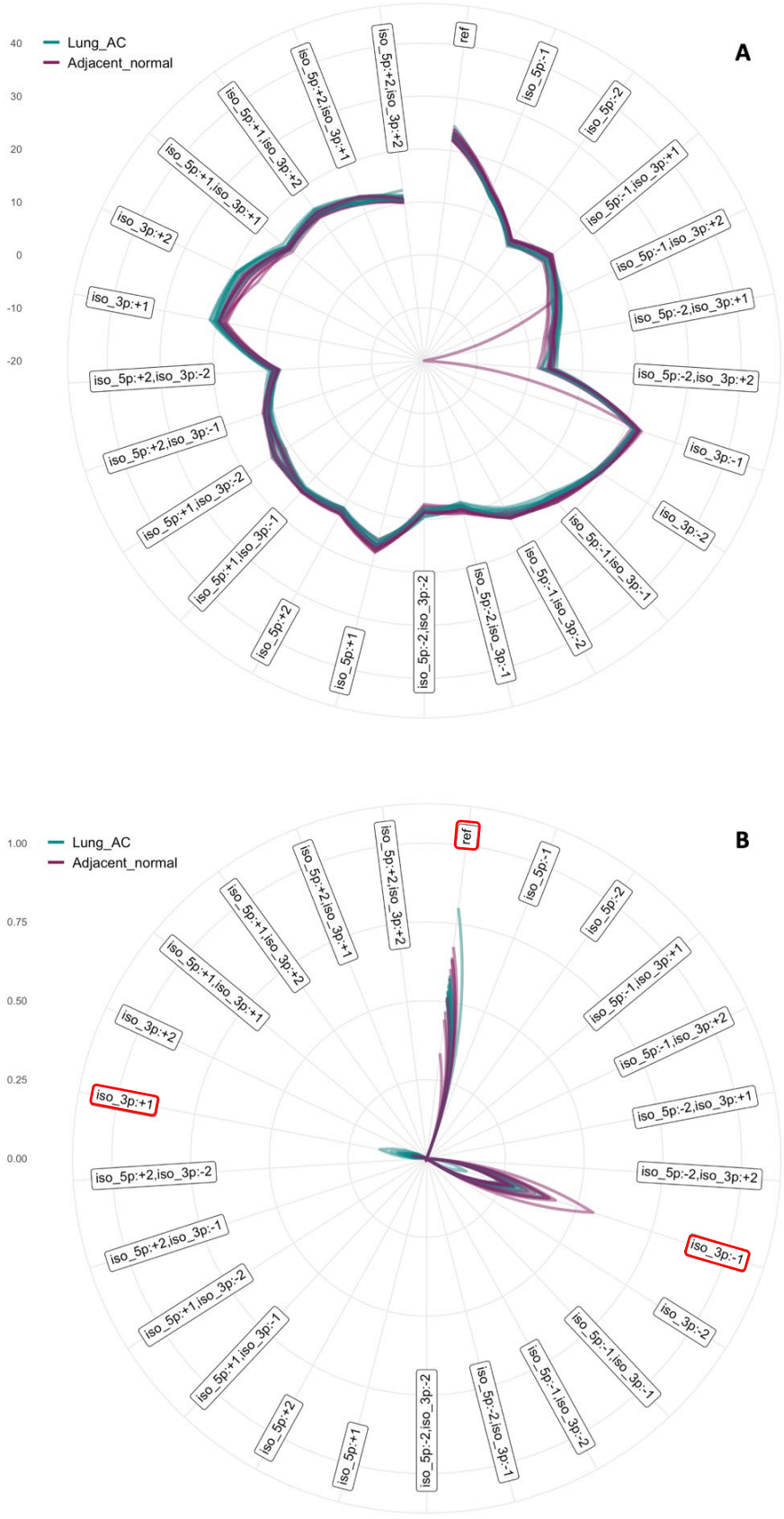


Figure 13. (a) Log2-expression and (b) composition of the isomiR populations across all miRNAs in LAC samples. The reference and single nucleotide 3' extension/deletion variants (in red) make up 92.9% and 93.6% of the isomiRs in disease state versus control, respectively.

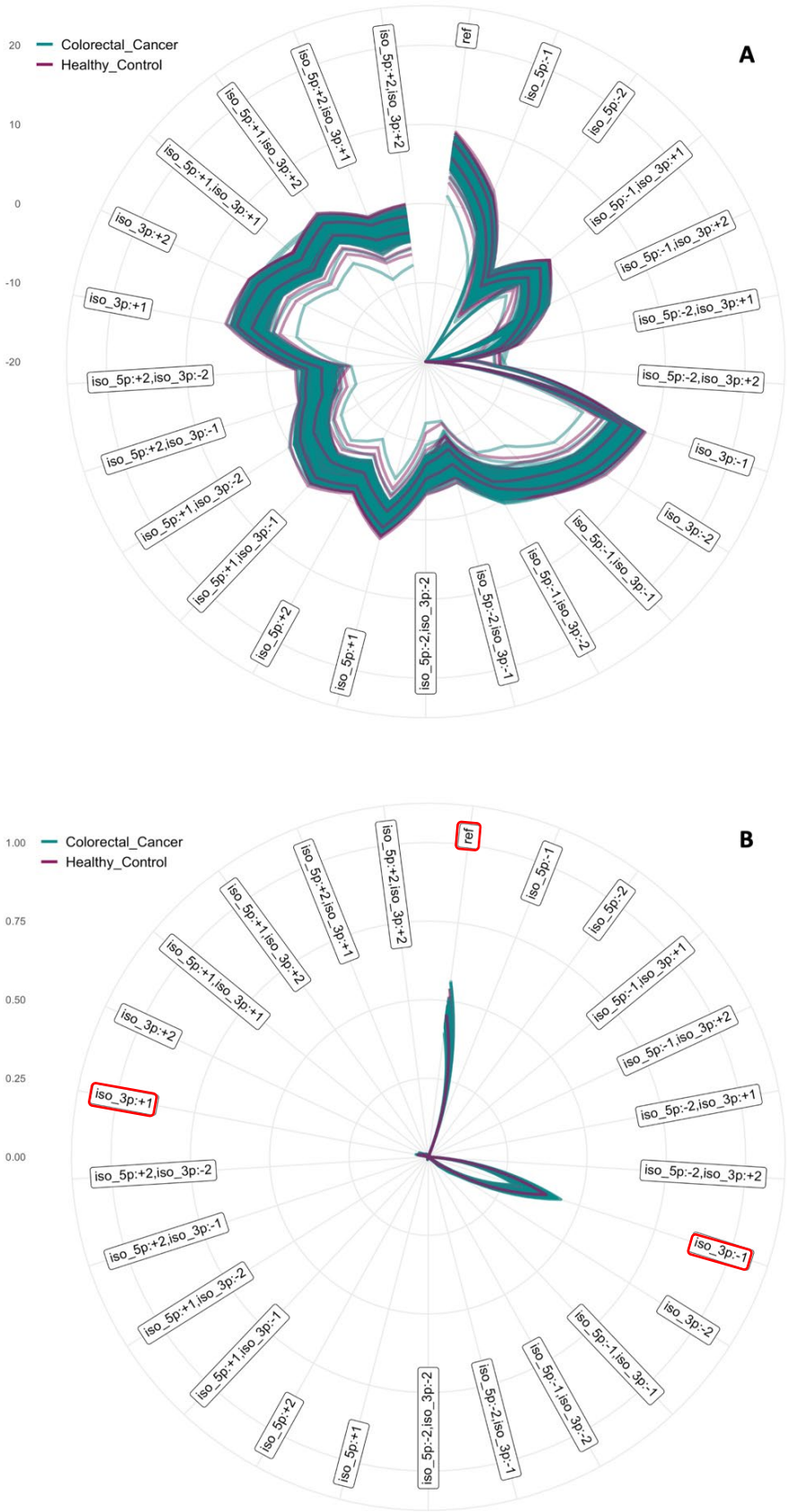


Figure 14. (a) Log2-expression and (b) composition of the isomiR populations across all miRNAs in CRC samples. The reference and single nucleotide 3' extension/deletion variants (in red) make up 86.4% and 86.6% of the isomiRs in disease state versus control, respectively.

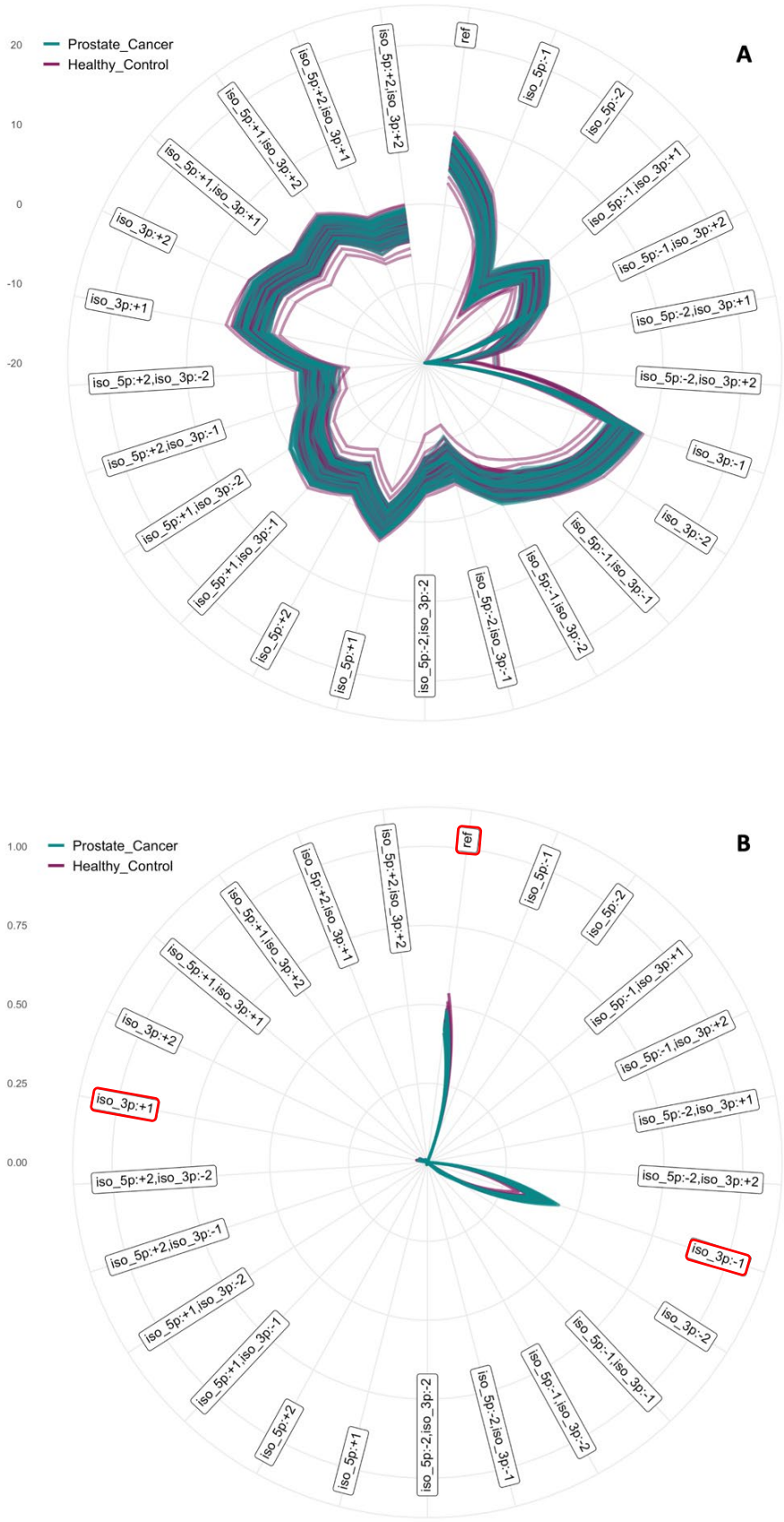


Figure 15. (a) Log2-expression and (b) composition of the isomiR populations across all miRNAs in PC samples. The reference and single nucleotide 3' extension/deletion variants (in red) make up 87.1% and 86.6% of the isomiRs in disease state versus control, respectively.

3.3.1 Aberrant isomiR expression levels

The differential expression analysis in section 3.2 above shows the presence and extent of differentially expressed isomiRs. As an example, we will look at `miR-21-5p`. `miR-21-5p` was the most frequently differentially expressed miRNA in both GBM and LAC samples, and has been reported to be a prognostic biomarker in at least 29 different diseases [48].

Figure 16 shows that `miR-21-5p` is, overall, more highly expressed in the disease state versus healthy state in both brain- and lung tissue. 14 and 19 out of 25 possible variants were differentially expressed in GBM and LAC samples, respectively. Variants having two-nucleotide deletions or extensions on the 5' end show both the lowest expression values and the lowest rate of differential expression. In PC samples, all four 3' trimming variants showed differential expression (Figure 19), but none were differentially expressed in CRC (Figure 18).

From Figure 17, the `miR-21-5p` isomiR profile appears to be fairly similar between GBM and LAC samples. The reference variant is the dominating variant, followed by 3' single nucleotide extension- and deletion events. This observation contrasts with the general trend of 3' deletions being more common than 3' extensions (Table 6). Variants `ref`, `3p:-1` and `3p:+1` collectively accounted for 80% of the `miR-21-5p` population in GBM samples, and 91% in LAC samples (Table 7).

In CRC and PC samples, the three top variants remain the same as for GBM and LAC samples, but the composition changes considerably. Abundance of the reference variant decreases, and that of `3p:-1` increases to make up over 50% of the total population.

Table 7. Average % of the population accounted for by the main variants of `miR-21-5p`.

	`ref`	`3p:-1`	`3p:+1`	**total**
GBM/control	47%/46%	8%/13%	25%/27%	80%/86%
LAC/control	56%/54%	11%/13%	24%/22%	91%/89%
CRC/control	34%/35%	53%/52%	11%/11%	86.4%/86.6%
PC/control	35%/35%	56%/52%	6.7%/11%	87.1%/86.6%

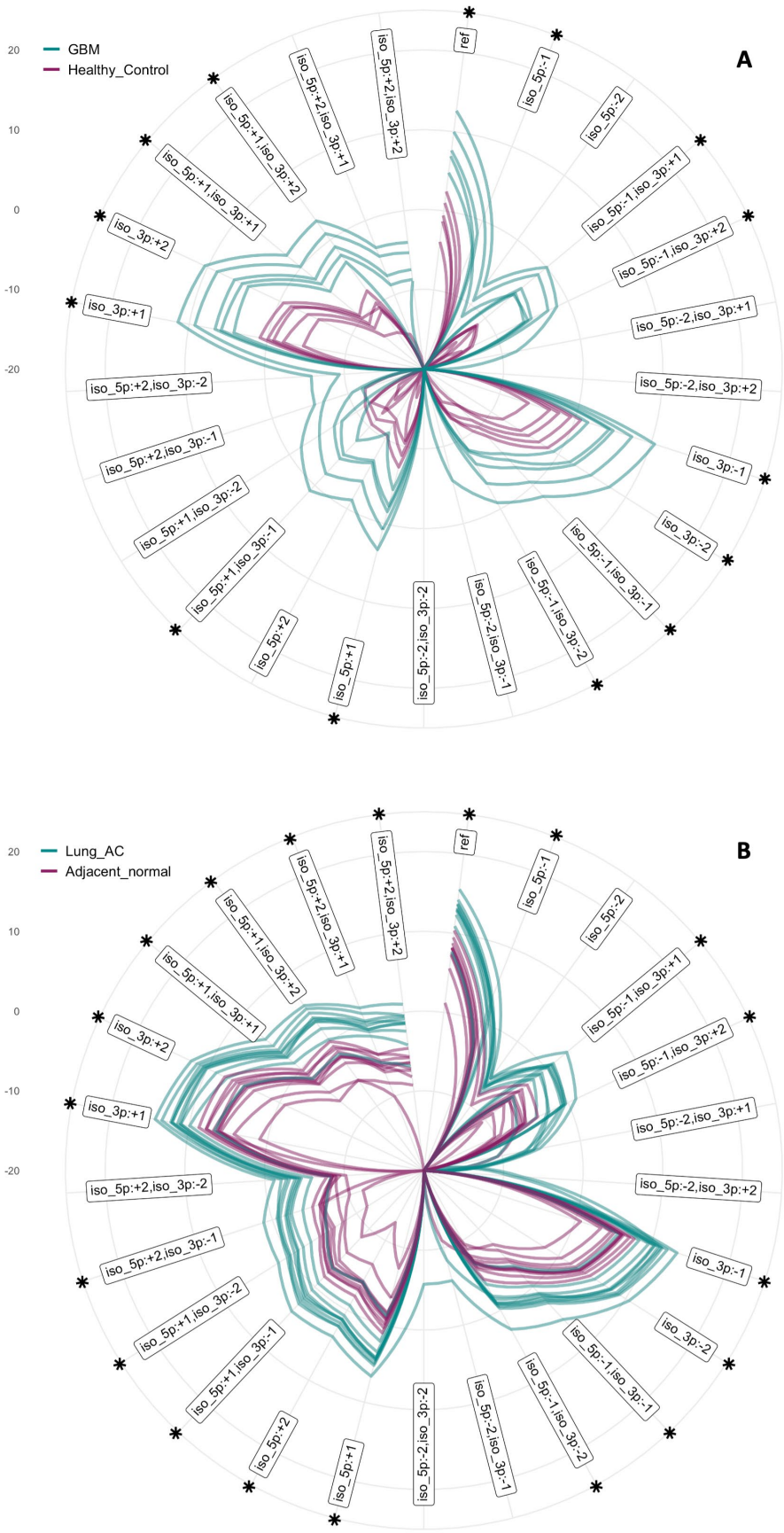


Figure 16. Log2-expression of the `miR-21-5p` isomiR population in (a) GBM samples and (b) LAC samples. Differentially expressed variants are marked with ✱.

Figure 17. Composition of the `miR-21-5p` isomiR population in (a) GBM samples and (b) LAC samples. In both conditions and tissue types, the dominant variants are the reference miRNA, and single nucleotide deletion or extensions on the 3' end (in red).

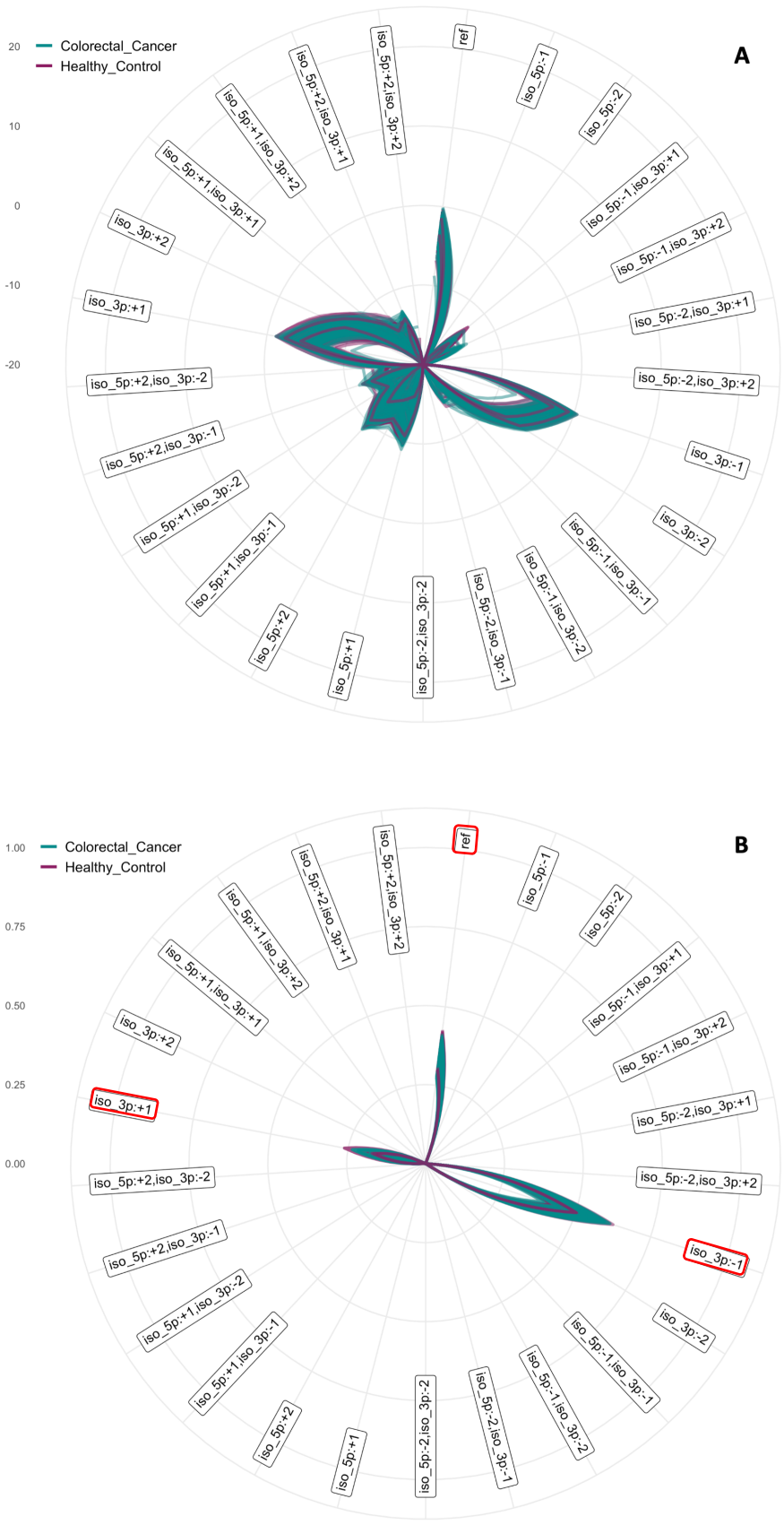


Figure 18. (a) Expression and (b) composition of the `miR-21-5p` isomiR population in CRC samples. The dominant variants are the reference miRNA, and single nucleotide deletion or extensions on the 3' end (in red). No variants were differentially expressed.

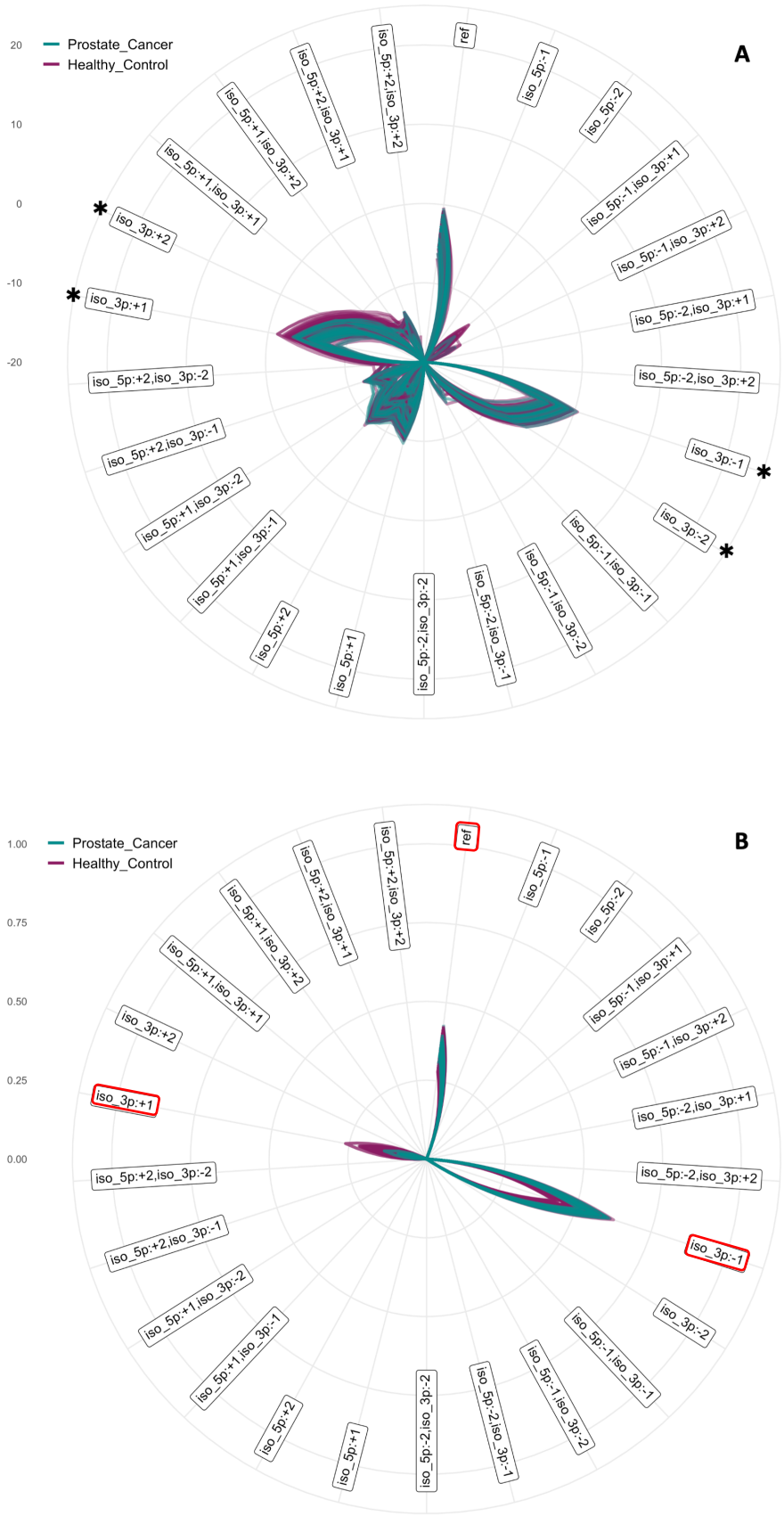


Figure 19. (a) Expression and (b) composition of the `miR-21-5p` isomiR population PC samples. The dominant variants are the reference miRNA, and single nucleotide deletion or extensions on the 3' end (in red). Differentially expressed variants are marked with ✱.

3.3.2 Divergent isomiR profiles

As seen in the previous sections, the composition of isomiR populations normally favours the reference variant, along with single nucleotide 3' deletion- and extensions. For individual miRNAs however, the composition of the population can vary considerably.

In CRC and PC samples, the miRNA with the highest number of differentially expressed variants was `miR-760` (Figure 11). For this isomiR population, the most commonly occurring variant was `3p:+1`, followed by `5p:-1,3p:+1`. The latter corresponds to a single nucleotide extension on both the 5' and 3' end, and for `miR-760`, this variant is more highly expressed than the reference (Figure 20 and Figure 21). Although the composition is atypical, it remains similar between the healthy and disease state in CRC and PC (Table 8).

The expression levels of `miR-760` differs between tissue types. Figure 22 and Figure 23 show expression levels and compositions of `miR-760` in GBM and LAC samples. Expression levels in these tissues and malignancies were markedly lower, and the isomiR profiles quite different to what is seen in CRC and PC samples. For GBM and LAC samples, the reference variant was much more abundant than for CRC and PC samples, where the `3p:+1` variant is favoured.

Only the `3p:+1` variant was significantly differentially expressed in GBM samples, while none of the variants were found differentially expressed in LAC.

Table 8. Average % of the population accounted for by the main variants of `miR-760`.

	`ref`	`3p:-1`	`3p:+1`	`5p:-1,3p:+1`	**total**
GBM/control	22%/19%	2.2%/3.8%	22%/36%	17%/17%	63.2%/75.8%
LAC/control	23%/16%	6.3%/1.4%	31%/29%	14%/12%	74.3%/58.4%
CRC/control	9%/9%	9%/10%	40%/39%	15%/15%	73%/73%
PC/control	9%/9%	10%/10%	40%/39%	14%/15%	73%/73%

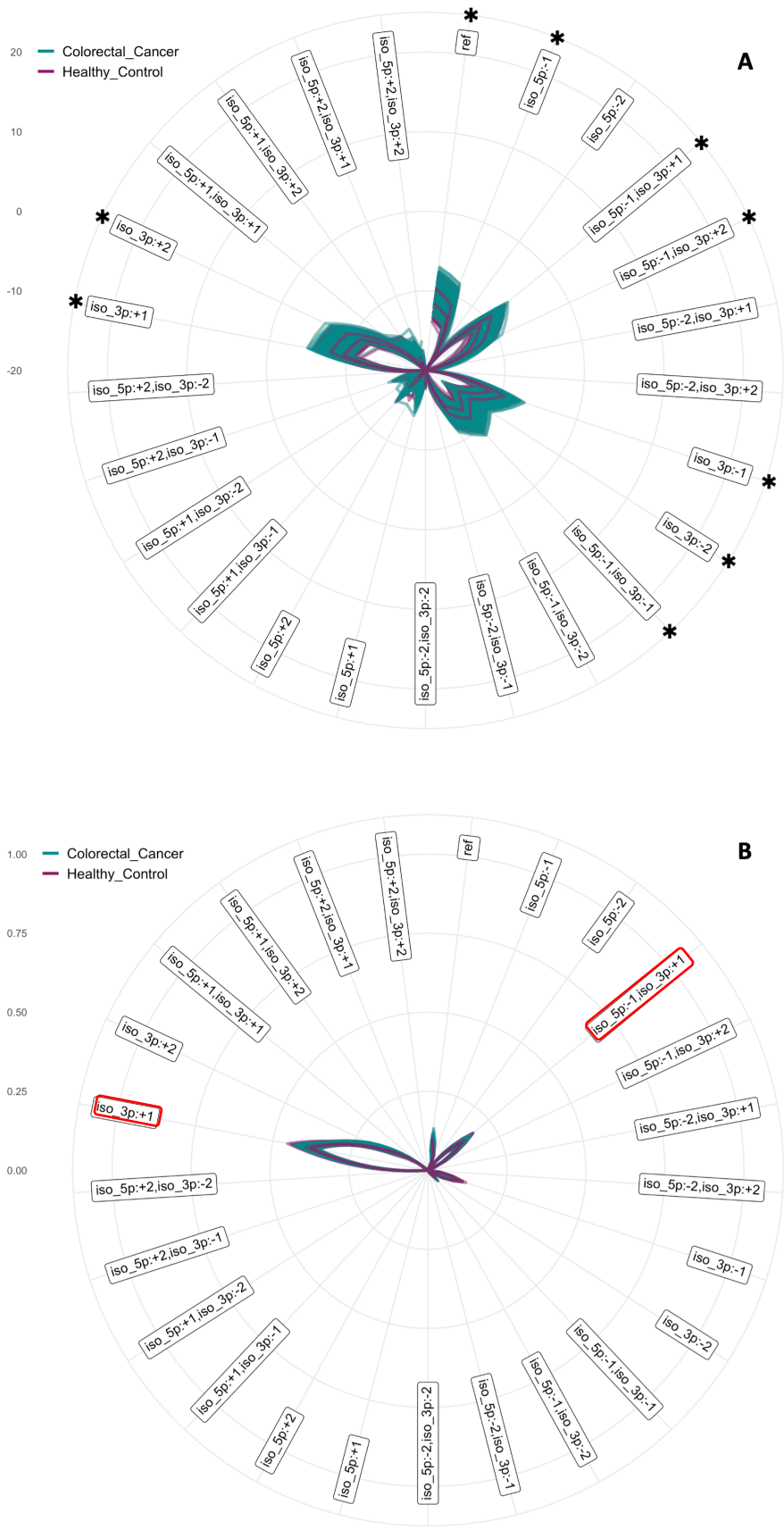


Figure 20. (a) Expression and (b) composition of the miR-760 isomiR population in CRC samples. Single nucleotide 3' extensions and a variant with both 3' and 5' extensions (in red) occur more frequently than the reference miRNA.

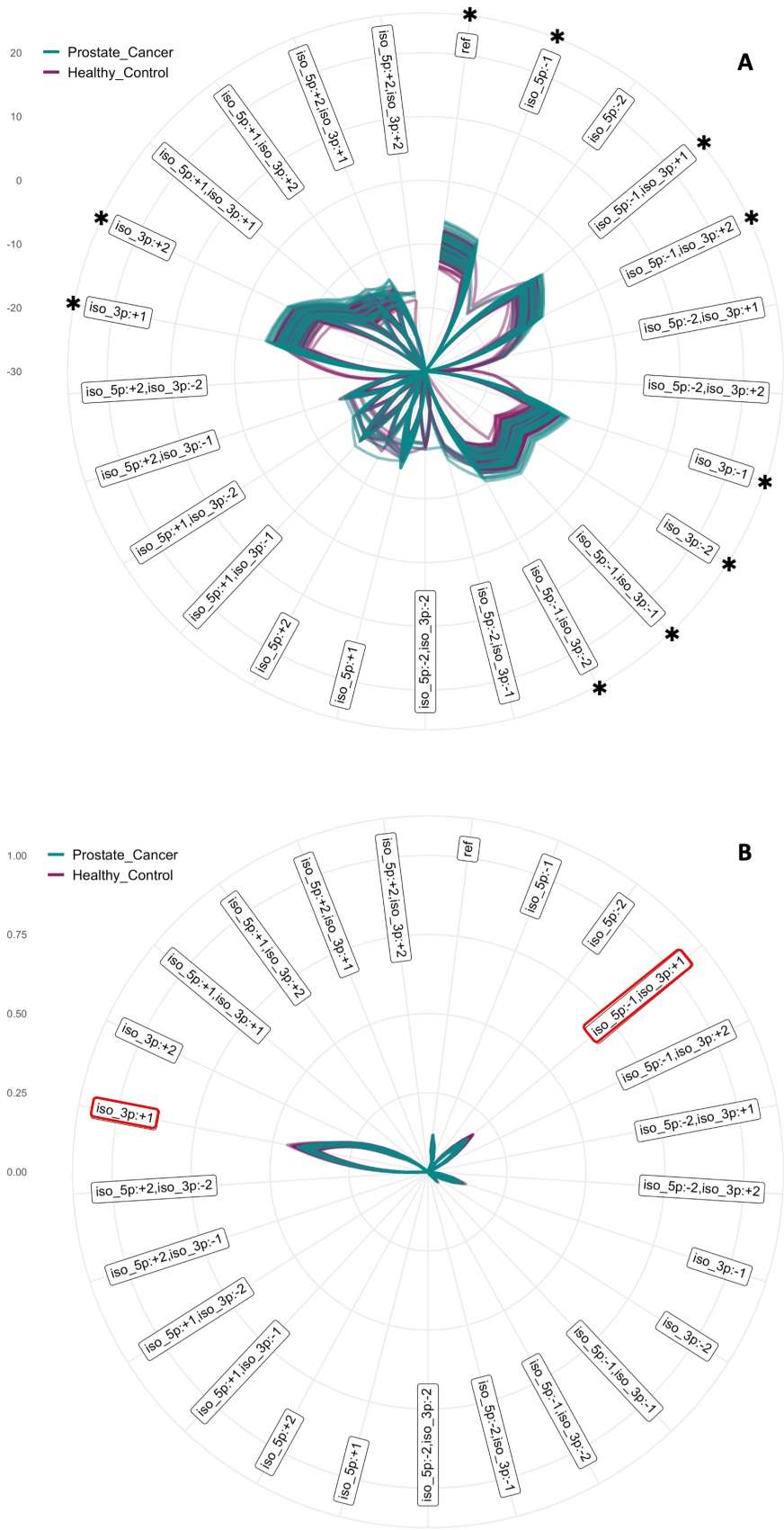


Figure 21. (a) Expression and (b) composition of the `miR-760` isomiR population in PC samples. Single nucleotide 3' extensions and a variant with both 3' and 5' extensions (in red) occur more frequently than the reference miRNA.

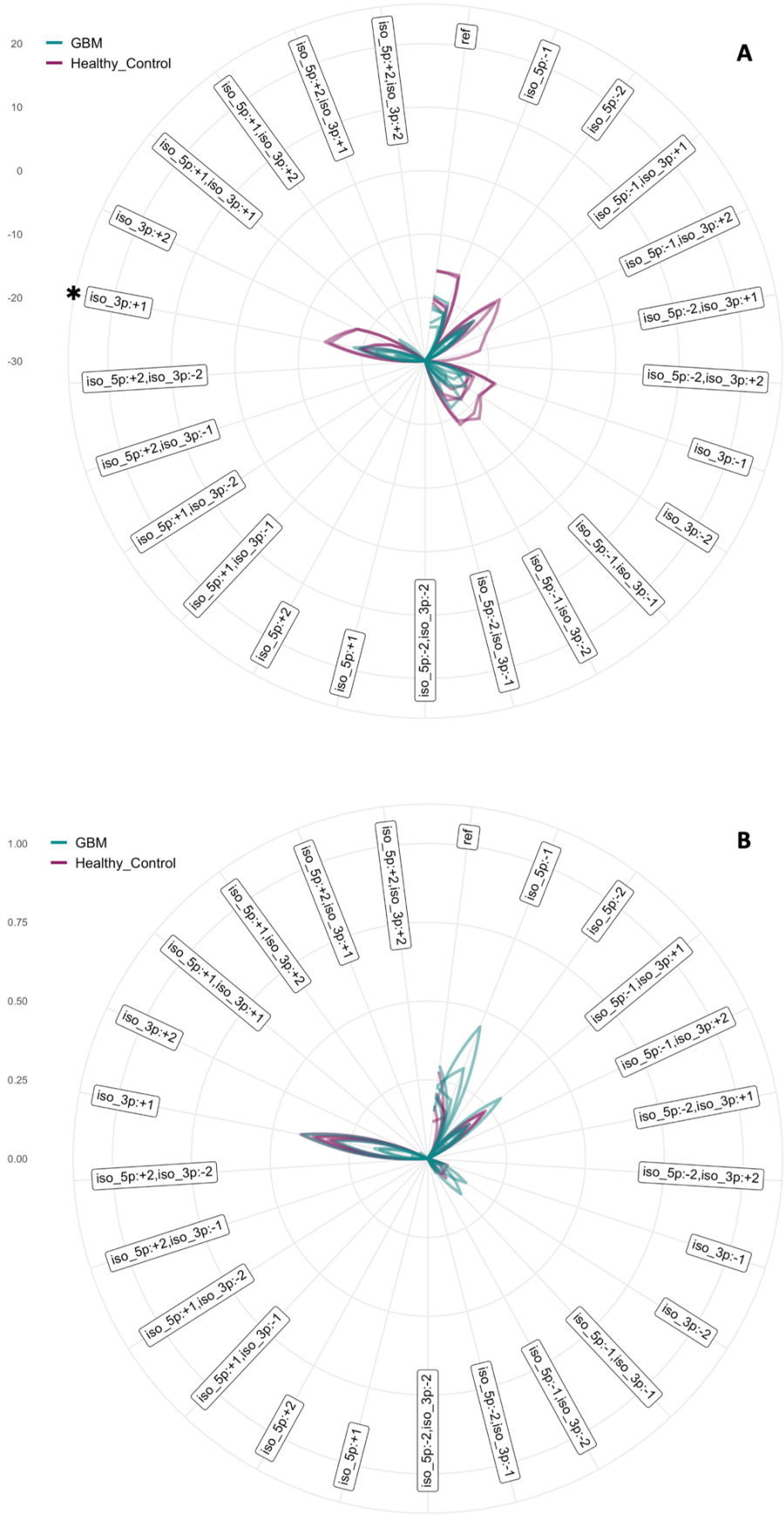


Figure 22. (a) Expression and (b) composition of the `miR-760` isomiR population in GBM samples. Only one variant was significantly differentially expressed (✱).

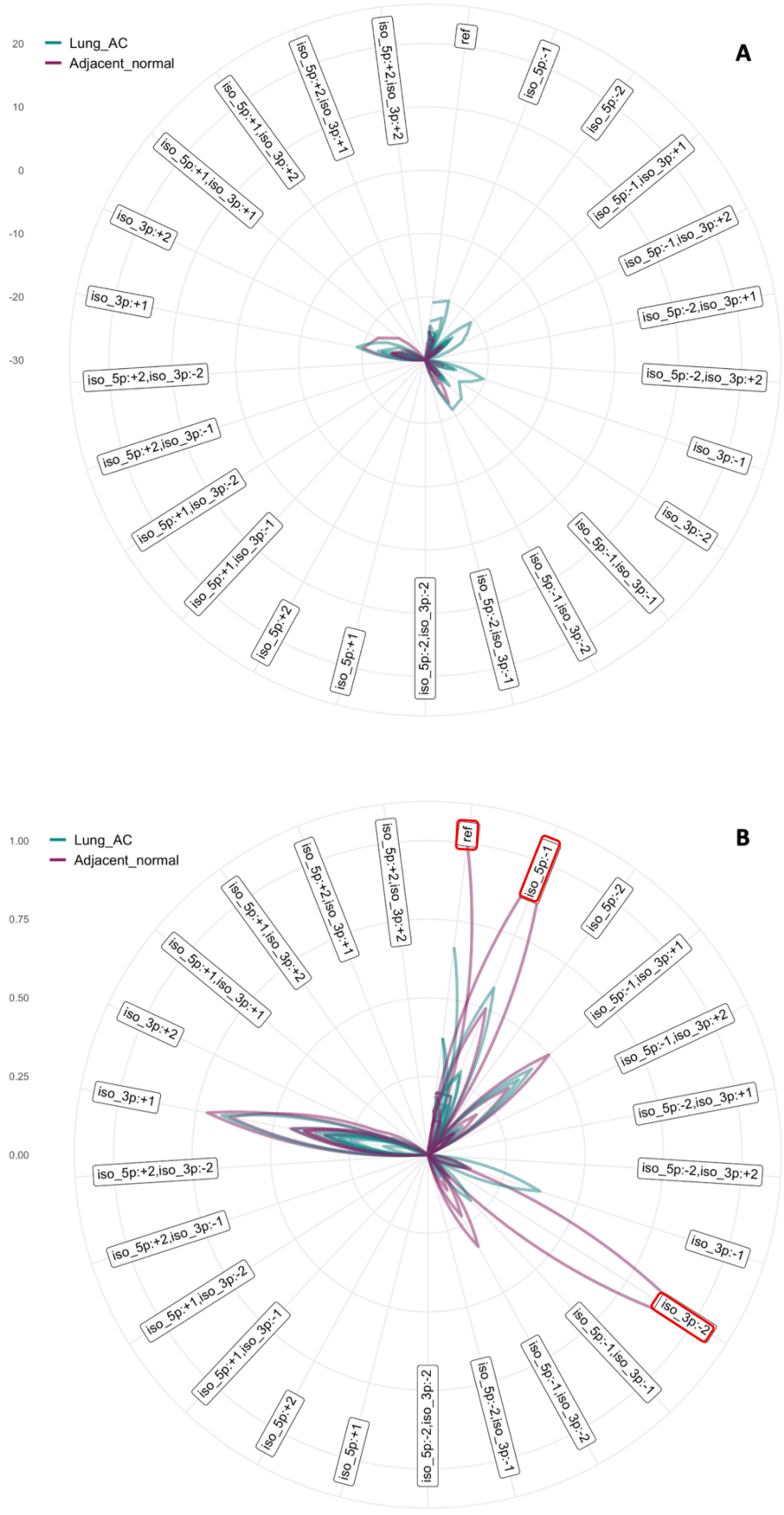


Figure 23. (a) Expression and (b) composition of the `miR-760` isomiR population in LAC samples. The extreme values seen in (b) (in red) correspond to 6 control samples where the given variant is the only one that is present.

3.4 Target prediction

Based on the results presented in section 3.3, target prediction was performed for the `miR-21-5p` isomiR population using the `miRAW` miRNA target prediction tool. The predictions show that the target network changes with the inclusion of multiple isomiRs.

For instance, target predictions of `miR-21-5p` and its isomiRs reveal unique gene targets for several variants (Figure 24). Predictions were done for 23 isomiRs, producing over 10,000 hits. After filtering out less reliable predictions, 540 predicted interactions for 9 isomiRs remained, of which 5 showed unique targeting interactions. The reference miRNA had no unique targets. Single- and double 3' extensions (`3p:+1` and `3p:+2`) were predicted to have 19 and 25 unique target mRNAs, respectively.

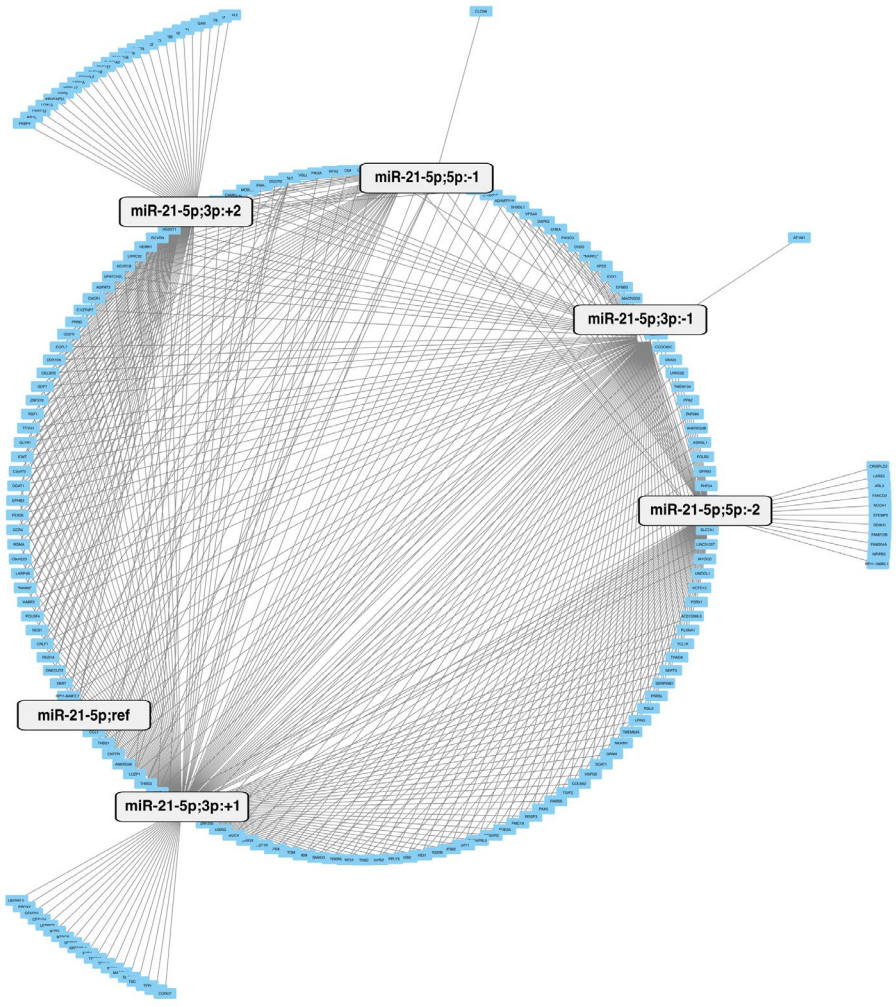


Figure 24. MMRN of `miR-21-5p` and its isomiRs. The nodes are isomiRs (grey labels), targets are mRNAs (blue labels), and edges are predicted miRNA-mRNA interactions. The inner circle represents mRNA targets shared between two or more variants. Labels outside of the circle are unique to the isomiR they are connected to.

4 Discussion

While the fundamental role of miRNAs in regulatory networks is widely studied, that role is certainly more complicated than once thought. Rather than existing as well-defined entities, miRNAs exist as a population of isomiRs. This population can vary, for example, across different tissue types and between healthy and disease study groups. While tools exist to study isomiRs, they are limited in terms of how they can investigate isomiR variation, and how they can identify important variants. The aim of this thesis was to develop methods for improved characterisation of miRNA populations expressed in NGS datasets that can be used to improve understanding of the roles of isomiRs in health and disease.

The methods have been implemented in an analysis pipeline that allows standardised and scalable analysis of raw NGS datasets. Thus, a user can specify a list of raw NGS data for a sample set and perform stepwise characterisation of isomiR populations that reports statistically significant changes in isomiR population between conditions, such as healthy and disease samples. Furthermore, the methods provide tools that allow the user to visually inspect observed isomiR populations in order to identify notable changes. However, the radar plots are generated to spot patterns and trends in large volumes of data, and not to be used independently for making claims of significance or for drawing conclusions.

The value of this approach has been demonstrated by investigating publicly available datasets to identify statistically significant changes in isomiR populations in various cancers.

4.1 Characterisation of isomiR populations

The results of this work support the view that, rather than being expressed as a single well-defined feature, miRNAs exist as populations of similar isomiRs. Up to 25 isoforms of a single miRNA could be detected by looking at trimming variants alone, with almost unlimited numbers of additional isoforms possible through the presence of polymorphisms. Our findings from investigating several sets of NGS datasets from various cancer studies indicate many different isoforms of a miRNA are present. Also, for individual miRNA populations, the expression levels of these alternative variants can be greater than that of the reference miRNA.

Such characteristics can be seen for `miR-760`, where the reference variant accounted for only ~10% of the population in the CRC and PC samples, and ~20% in the GBM and LAC samples. `miR-760` has previously been claimed to be downregulated in CRC [49], in contrast to the observations in this work. If standard analyses fail to capture isomiR variants, these miRNA levels could be massively underreported. Whether the inconsistent findings are due to the divergent isomiR profile of this miRNA or is a consequence of differences in the analyses is unclear. Regardless, it highlights the importance of investigating miRNA populations more thoroughly.

The contrasting isomiR profiles of `miR-760` to `miR-21-5p` also suggest that multiple mechanisms are responsible for isomiR variation. It is likely much more complicated than the term ‘trimming variant’ implies. If there was systematic, random error in Dicer cleavage, comparable characteristics would have been observed in all miRNAs.

Among tissues, conditions, and miRNAs, the most commonly occurring isomiRs were 3’ trimming variants. The argument could be made that they are not of functional importance as they do not impair canonical seed region pairing. However, target predictions for the `miR-21-5p` isomiR population indicate that the 3’ variants introduce many additional and unique targeting events, suggesting that they have a relevant and potentially disruptive effect on the regulatory mRNA network (Figure 24). The 3’ extension events (`3p:+1` and `3p:+2`) were predicted to have 19 and 25 unique target genes, respectively. The functional ramifications of these additional targets have not been explored as part of this work. Nonetheless, whether they have alternative functional roles or not, they can have an effect merely by contributing to increased expression levels of the miRNA (i.e., failing to take them into consideration can lead to undercounting of miRNA reads, as was observed for `miR-760`).

A potential explanation for the lack of 5’ isomiRs is the canonical targeting associated with seed region binding between the miRNA and its target. This represents a core functionality and additional changes at the 5’ end of the miRNA would cause a frame shift-like event changing the seed region complementarity (Figure 5). This would introduce much greater disruption to regulatory function, i.e., there is a form of selective pressure at work in the isomiR generation process.

A further observation from the study of the publicly available cancer datasets is that isomiR profiles seem to display condition- and tissue specificity, much like miRNA expression levels [50].

In these datasets, condition specificity was marked by:

- The reference variant decreased and 3' variants increased in GBM samples compared to controls (Table 6)
- 3' extensions increased and 3' deletions decreased in LAC samples compared to controls (Table 6)

Similarly, tissue specificity was shown by:

- The reference variant is much less abundant in microvesicles, and most abundant in brain tissue (Table 6)
- In `miR-21-5p`, isomiR composition changes from favouring 3' deletions in microvesicles, to 3' extensions in brain- and lung tissue (Table 7)
- The reference variant of `miR-760` is decreased in microvesicles compared to brain- and lung tissue (Table 8)

In the studied datasets, samples sourced from microvesicles show isomiR patterns that are strikingly different from the tissue samples. Also, isomiR profiles in microvesicles remained very similar between healthy and disease states, compared to tissue samples. This warrants further investigation as no metric has been developed for comparing isomiR profiles at this point. Divergences in miRNA content between microvesicles and tissues have previously been described [51], and direct comparisons between the two groups should therefore be done with caution. Still, it relays interesting information on exosomal miRNA distributions and how they differ from those in tissues. Selective forces seem to choose subpopulations of miRNAs to load into microvesicles, which might extend to the preference of selected variants as well.

Microvesicles also generally display less disruptions to miRNA expression levels. From the differential expression analysis results, log2fold-changes are notably smaller in CRC and PC samples (Figure 11) than in GBM and LAC samples (Figure 10). They appear to reach statistical significance due to the larger sample sizes (n=200 and n=72, vs n=5 and n=10, respectively). The isomiR changes reportedly unique to CRC and PC samples might also be present in the other conditions, but sample sizes are too small for them to reach significance.

Regardless of the small sample size, over 300 isomiRs were uniquely differentially expressed in GBM samples. The markedly high numbers could be a consequence of the different cell types in the GBM tumour and the frontal cortex of the control samples. The former is a neoplasm derived from glial cells, whereas frontal cortex is also rich in neural bodies, which presumably have different miRNA expression profiles. This might be a confounding factor and should be kept in mind.

4.2 Software optimisation

The methods in this thesis were developed with the four FAIR-principles in mind – findable, accessible, interoperable, and reusable. However, additional refinements can be made to the software to improve user experience.

First of all, it would be preferable to slightly simplify the variant classification used in the `mirGFF3` class. A proposed modification is to avoid the compound events described in the original format in which deletions can be followed by non-template extensions. Figure 25 shows a suggested simplification where the "deletion with non-template extension" event is removed. This would instead be termed an SNV and classified according to the nomenclature in Table 3. We do not know the mechanism(s) responsible for these events, and thus the general term SNV could be more appropriate. In addition to reducing confusion, this modification would reduce the number of methods needed for the classification, which increases code readability, and reduces the maintenance load.

Secondly, the post-processing scripts can be optimised in terms of user-friendliness and automation. The intention is to further develop the Python data-wrangling script to allow it to run as a command-line program, as well as streamlining the R scripts for processing larger amounts of data simultaneously.

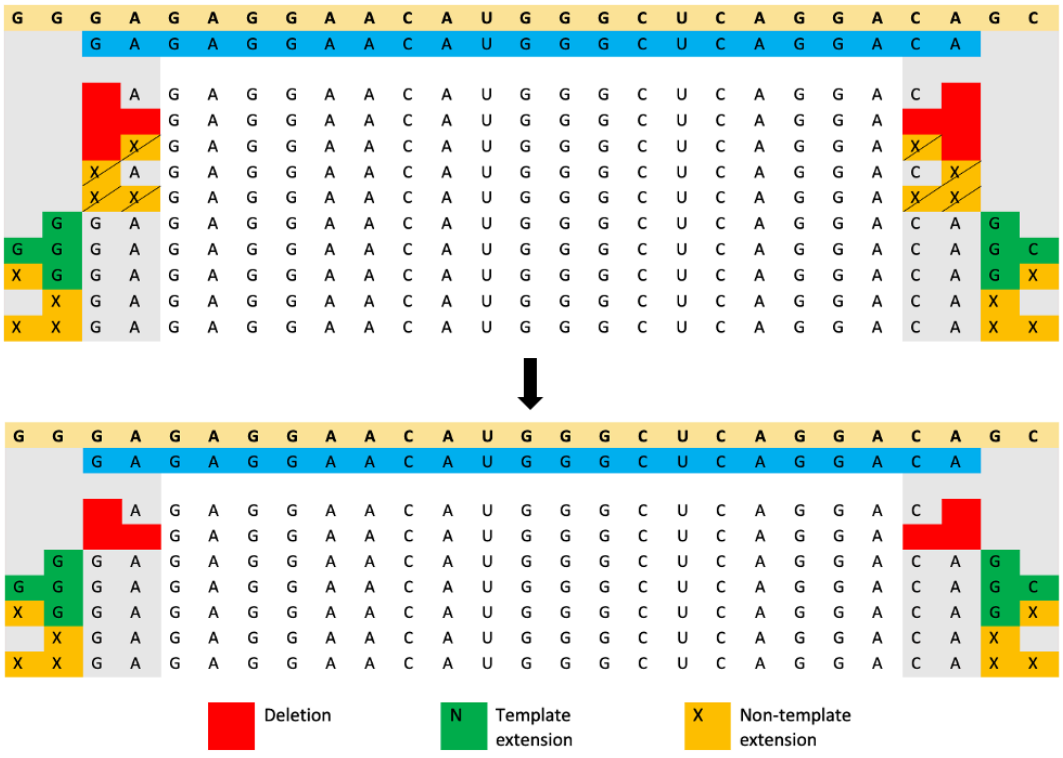


Figure 25. Suggested simplification of the mirGFF3 5'/3' isomiR classification.

4.3 Further directions

The methods developed in this project were developed to facilitate standardised miRNA profiling based on Small RNA-Seq data. Based on our investigation of publicly available datasets, the following opportunities include, but are not limited to:

1) Investigation of SNV-isomiRs

2) Systematic investigation of alterations in isomiR profiles

3) Investigation of the targets and functional implications of isomiRs

4) Investigation of the possible mechanisms underlying the production of isomiR variants

References

1. Crick, F.H. *On protein synthesis*. in *Symp Soc Exp Biol*. 1958.
2. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-563.
3. *Created with https://biorender.com*.
4. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
5. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proceedings of the National Academy of Sciences, 2009. **106**(23): p. 9362-9367.
6. Freund, M.K., et al., *Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits.* American journal of human genetics, 2018. **103**(4): p. 535-552.
7. Rheinbay, E., et al., *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.* Nature, 2020. **578**(7793): p. 102-111.
8. *Adapted from "Types of RNA Produced in Cells", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates*.
9. Lange, M., R. Begolli, and A. Giakountis, *Non-Coding Variants in Cancer: Mechanistic Insights and Clinical Potential for Personalized Medicine.* Non-Coding RNA, 2021. **7**(3): p. 47.
10. Jungers, C.F. and S. Djuranovic, *Modulation of miRISC-Mediated Gene Silencing in Eukaryotes.* Frontiers in molecular biosciences, 2022. **9**: p. 832916-832916.
11. Walavalkar, K. and D. Notani, *Beyond the coding genome: non-coding mutations and cancer.* Frontiers in bioscience (Landmark edition), 2020. **25**: p. 1828-1838.
12. Zhang, L., Q. Lu, and C. Chang, *Epigenetics in Health and Disease*, in *Epigenetics in Allergy and Autoimmunity*, C. Chang and Q. Lu, Editors. 2020, Springer Singapore: Singapore. p. 3-55.
13. Ferreira, H.J. and M. Esteller, *Non-coding RNAs, epigenetics, and cancer: tying it all together.* Cancer and Metastasis Reviews, 2018. **37**(1): p. 55-73.
14. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
15. Krol, J., I. Loedige, and W. Filipowicz, *The widespread regulation of microRNA biogenesis, function and decay.* Nature Reviews Genetics, 2010. **11**(9): p. 597-610.
16. Incorvaia, L., et al., *A "Lymphocyte MicroRNA Signature" as Predictive Biomarker of Immunotherapy Response and Plasma PD-1/PD-L1 Expression Levels in Patients with Metastatic Renal Cell Carcinoma: Pointing towards Epigenetic Reprogramming.* Cancers, 2020. **12**(11): p. 3396.
17. Janssen, H.L., et al., *Treatment of HCV infection by targeting microRNA.* N Engl J Med, 2013. **368**(18): p. 1685-94.
18. Jin, W., et al., *Structural Basis for pri-miRNA Recognition by Drosha.* Molecular Cell, 2020. **78**(3): p. 423-433.e5.
19. Rorbach, G., O. Unold, and B.M. Konopka, *Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods.* Scientific Reports, 2018. **8**(1): p. 7560.
20. Cheloufi, S., et al., *A dicer-independent miRNA biogenesis pathway that requires Ago catalysis.* Nature, 2010. **465**(7298): p. 584-589.
21. Ratti, M., et al., *MicroRNAs (miRNAs) and Long Non-Coding RNAs (lncRNAs) as New Tools for Cancer Therapy: First Steps from Bench to Bedside.* Targeted oncology, 2020. **15**(3): p. 261-278.
22. Nawalpuri, B., S. Ravindran, and R.S. Muddashetty, *The Role of Dynamic miRISC During Neuronal Development.* Frontiers in Molecular Biosciences, 2020. **7**.
23. Xu, K., et al., *MicroRNA-mediated target mRNA cleavage and 3'-uridylation in human cells.* Scientific Reports, 2016. **6**(1): p. 30242.
24. Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones, *miRBase: from microRNA sequences to function.* Nucleic Acids Research, 2018. **47**(D1): p. D155-D162.

25. Morin, R.D., et al., *Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.* Genome research, 2008. **18**(4): p. 610-621.
26. Tan, G.C., et al., *5′ isomiR variation is of functional and evolutionary importance.* Nucleic Acids Research, 2014. **42**(14): p. 9424-9435.
27. Tan, G.C. and N. Dibb, *IsomiRs have functional importance.* Malays J Pathol, 2015. **37**(2): p. 73-81.
28. Cloonan, N., et al., *MicroRNAs and their isomiRs function cooperatively to target common biological pathways.* Genome Biology, 2011. **12**(12): p. R126.
29. Llorens, F., et al., *A highly expressed miR-101 isomiR is a functional silencing small RNA.* BMC Genomics, 2013. **14**(1): p. 104.
30. Desvignes, T., et al., *Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API.* Bioinformatics, 2019. **36**(3): p. 698-703.
31. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era — concepts and misconceptions.* Nature Reviews Genetics, 2008. **9**(4): p. 255-266.
32. Génin, E., *Missing heritability of complex diseases: case solved?* Human Genetics, 2020. **139**(1): p. 103-113.
33. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.
34. *Adapted from "The Principle of a Genome-wide Association Study (GWAS)", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.*
35. Telonis, A.G., et al., *Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types.* Nucleic Acids Research, 2017. **45**(6): p. 2973-2985.
36. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 2016. **3**(1): p. 160018.
37. Fakhoury, S., et al. *The Effect of Poor Source Code Lexicon and Readability on Developers' Cognitive Load*. in *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*. 2018.
38. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
39. *FastQC*. 2015.
40. *Adapted from "Next Generation Sequencing Data Processing", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates.*
41. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
42. Pliatsika, V., et al., *MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments.* Bioinformatics, 2016. **32**(16): p. 2481-9.
43. *SRA Toolkit Development Team*. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
44. *pheatmap (RRID:SCR_016418)*. Available from: https://www.rdocumentation.org/packages/pheatmap/versions/0.2/topics/pheatmap.
45. Wickham, H., *Ggplot2: Elegant graphics for data analysis*. 2 ed. Use R! 2016, Cham, Switzerland: Springer International Publishing. 260.
46. Pla, A., X. Zhong, and S. Rayner, *miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts.* PLOS Computational Biology, 2018. **14**(7): p. e1006185.
47. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
48. Jenike, A.E. and M.K. Halushka, *miR-21: a non-specific biomarker of all maladies.* Biomarker Research, 2021. **9**(1): p. 18.
49. Manvati, M.K.S., et al., *Association of miR-760 with cancer: An overview.* Gene, 2020. **747**: p. 144648.
50. Ludwig, N., et al., *Distribution of miRNA expression across human tissues.* Nucleic Acids Research, 2016. **44**(8): p. 3865-3877.
51. Chen, M., et al., *Distinct shed microvesicle and exosome microRNA signatures reveal diagnostic markers for colorectal cancer.* PLOS ONE, 2019. **14**(1): p. e0210003.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway