



UNIVERSITÀ
DEGLI STUDI
FIRENZE

UNIVERSITÀ DEGLI STUDI DI FIRENZE
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE (DINFO)
CORSO DI DOTTORATO IN INGEGNERIA DELL'INFORMAZIONE
CURRICULUM: DINAMICA NON LINEARE E SISTEMI COMPLESSI

DEVELOPMENT OF
COMPUTATIONAL METHODS FOR
LONG READ SEQUENCING DATA
ANALYSIS IN THE OMICS ERA

Candidate

Marta Baragli

Supervisors

Prof. Alberto Magi

Prof. Leonardo Bocchi

PhD Coordinator

Prof. Stefano Ricci

CICLO XXXVIII, 2022-2025

Università degli Studi di Firenze, Dipartimento di Ingegneria
dell'Informazione (DINFO).

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Engineering. Copyright © 2026 by
Marta Baragli.

Contents

Contents	iii
1 Introduction	1
1.1 The Omic Sciences	1
1.1.1 Genomics	1
1.1.2 Epigenomics	3
1.1.3 Transcriptomics	5
1.2 DNA sequencing evolution	6
1.2.1 First-Generation Sequencing	6
1.2.2 Massively Parallel Sequencing	7
1.2.3 Long-Read Sequencing	9
2 Aim of the Thesis	13
3 Structural Variant detection	17
3.1 Background	18
3.1.1 Cytogenetics and Hybridization-based techniques . . .	18
3.1.2 NGS-based SV detection	19
3.1.3 TGS-based SV detection	20
3.2 State of the art	22
3.3 GASOLINE	25
3.3.1 Workflow	26
3.3.2 Somatic SV calling	27
3.4 Results	28
3.4.1 Germline variants	28
3.4.2 Somatic variants	31
3.5 Materials and Methods	36
3.5.1 Simulation of synthetic datasets with PBSIM2	36

3.5.2	Comparison with existing SV detection tools	36
3.5.3	NA24385 dataset (GIAB)	37
3.5.4	COLO829 dataset	38
3.6	Conclusions	38
3.6.1	Future developments	40
4	Differential Methylation detection	41
4.1	Background	42
4.1.1	Early assays for methylation detection	42
4.1.2	Bisulfite-based methylation sequencing	43
4.1.3	Methylation Arrays	44
4.1.4	Direct long-read sequencing	45
4.2	State of the art	46
4.3	Results	49
4.3.1	Methylation Entropy	49
4.3.2	BiSLM	53
4.3.3	Annotation	58
4.3.4	Acute Myeloid Leukemia	59
4.3.5	Human Peripheral Glial Cells	64
4.4	Materials and Methods	69
4.4.1	AML samples: sequencing and data preparation . . .	69
4.4.2	HPGC samples: sequencing and data preparation . . .	69
4.4.3	Over-Representation Analysis	70
4.5	Conclusion	70
4.5.1	Future developments	71
5	Epitranscriptomics	73
5.1	Introduction	74
5.1.1	RNA Editing and the Epitranscriptome	74
5.1.2	Evolution of Detection Methodologies	76
5.2	State of the Art	79
5.2.1	Resquigging	80
5.3	Results	82
5.3.1	Signal Extraction and Realignment Assessment	82
5.3.2	Self-Alignment	86
5.3.3	K-mer based aggregation	88
5.4	Materials and Methods	91
5.4.1	Data Preparation and Processing	91

5.4.2	Signal Extraction and Realignment Pipelines	92
5.4.3	Custom Signal Extraction Pipelines	93
5.4.4	Statistical Analysis and Performance Metrics	94
5.5	Conclusion	95
6	Conclusion	99
6.1	Summary of contribution	99
A	Publications	103

Chapter 1

Introduction

1.1 The Omic Sciences

In the last few decades, the advent of increasingly accurate, reliable, and cost-effective technologies for biological analyses has revolutionized our ability to investigate complex biological systems. This shift has led to the emergence of the omic sciences, a family of disciplines, including genomics, epigenomics, proteomics and transcriptomics, that aim to study entire classes of molecules (DNA, RNA, proteins) or molecular features (regulatory states) rather than isolated elements. This systems-level perspective has transformed modern biology, moving beyond the reductionist paradigm to a holistic understanding of how molecular layers interact to orchestrate the dynamic organization of living systems.

1.1.1 Genomics

Genomics is the branch of the omic sciences concerned with the study of the entire set of DNA molecules within an organism, not only in terms of sequence but also from a structural and functional point of view. Specifically, the subject of genomic studies is not only related to coding DNA sequences, but also non-coding regions, as well as regulatory elements that modulate gene activity and the interactions among them. From this perspective, genomics seeks to understand how alterations in the genome shape phenotypes and contribute to both physiological diversity and disease.

Genomic variability refers to the spectrum of differences that can occur

within the DNA sequence of individuals or populations, spanning from single nucleotide changes to large chromosomal rearrangements. Broadly, these variations are categorized into small-scale and large-scale events. The former includes single nucleotide variants (SNVs) and small insertions or deletions (InDels), while the latter comprises structural variants (SVs), which affect regions larger than 50 base pairs and include deletions, duplications, insertions, inversions, and translocations, as well as more complex rearrangements. Together, these forms of variation contribute to the unique genomic landscape of every individual.

From a physiological standpoint, genomic variability is the foundation of biological diversity and evolutionary adaptation. Neutral variants, which do not affect gene function, contribute to population-level heterogeneity and serve as invaluable markers for studies on ancestry, demography, and evolution [1]. Other variants can directly modulate gene expression, protein structure, or regulatory interactions, generating phenotypic diversity in traits such as metabolism, height, or pigmentation.

Genomic variability also plays a crucial role in the onset and progression of disease. Highly penetrant rare mutations can directly cause phenotypes known as monogenic disorders, such as cystic fibrosis or sickle-cell anemia. In contrast, combinations of common variants that have little or no effect when isolated, can contribute to polygenic and multifactorial conditions, including diabetes, cardiovascular disease, and psychiatric disorders. Structural alterations can modify gene dosage, disrupt regulatory domains, or generate fusion transcripts, mechanisms that are particularly relevant in cancer genomics.

For these reasons, the comprehensive study of genomic variability has become one of the cornerstones of modern biomedical research. The advent of high-throughput DNA sequencing technologies has dramatically expanded our ability to detect and characterize the entire spectrum of genomic variants, leading to large-scale initiatives such as the 1000 Genomes Project [1]. NGS enabled the discovery of millions of SNVs and InDels across populations, paving the way for genome-wide association studies (GWAS) linking genetic variants to phenotypic traits. More recently, long-read sequencing technologies, such as those developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), have provided the means to resolve complex genomic regions with unprecedented accuracy [2], [3], revealing layers of variation that were previously inaccessible.

In summary, genomic variability constitutes both the blueprint of human diversity and a key driver of disease. Its characterization requires technologies capable of spanning the full spectrum of variant sizes and complexities, making DNA sequencing and methods for data analysis an indispensable tool for modern biology.

1.1.2 Epigenomics

Epigenomics refers to the comprehensive study of heritable molecular mechanisms that regulate gene activity without altering the underlying DNA sequence. These mechanisms include DNA methylation, histone modifications, chromatin remodeling, and the action of non-coding RNAs. Together, they define the epigenetic landscape of the cell, which determines its transcriptional identity and plasticity. Epigenomic regulation is essential for embryonic development, genomic imprinting, and cell differentiation, allowing genetically identical cells to acquire distinct functional identities. Moreover, epigenetic marks are dynamically regulated in response to environmental factors such as diet, aging, and stress, thereby integrating intrinsic and extrinsic cues into gene regulation [4].

Alterations in the epigenome play a major role in the pathogenesis of numerous diseases. In cancer, aberrant DNA methylation or histone modification patterns can lead to the silencing of tumor suppressor genes or activation of oncogenes [5]. Dysregulated chromatin states have also been associated with neurological, autoimmune, metabolic disorders and others [6]–[8]. Crucially, unlike genetic mutations, many epigenetic alterations are reversible, making them attractive targets for therapeutic intervention through so-called epigenetic drugs [9].

DNA methylation

DNA methylation is one of the most studied and biologically significant epigenetic modifications. It involves the covalent addition of a methyl group to the 5' position of cytosine residues, primarily within CpG dinucleotides, forming 5-methylcytosine (5mC) [10]. In mammals, CpG sites are generally sparse and methylated throughout the genome, except within CpG islands (CGIs), which are often located at gene promoters and remain unmethylated in normal conditions to allow transcriptional activation [11]. CpG shores

and shelves surrounding these islands exhibit more variable, tissue-specific methylation, contributing to regulatory diversity.

Functionally, DNA methylation exerts its effects primarily by influencing chromatin structure and DNA accessibility to transcription factors (TFs). Promoter methylation typically represses gene expression, whereas gene body methylation can promote transcriptional elongation and reduce spurious initiation [12]. Methylated CpGs are recognized by methyl-CpG-binding domain (MBD) proteins, which recruit chromatin remodelers and histone-modifying enzymes, establishing stable repressive states [13].

The establishment and maintenance of methylation patterns are tightly regulated. DNMT1 primarily maintains methylation fidelity during DNA replication, while DNMT3A and DNMT3B mediate de novo methylation, establishing new methylation marks during embryogenesis, cell differentiation, and cellular reprogramming [14]. Active DNA demethylation, is instead mainly catalyzed by the TET family of enzymes [15]. The dynamic balance between these opposing enzymatic systems allows cells to maintain epigenetic stability while preserving the plasticity required for developmental and environmental adaptation.

Physiologically, DNA methylation regulates several processes including embryonic development, genomic imprinting and X-chromosome inactivation, and is critical for maintaining cellular identity. Aberrant DNA methylation patterns are, however, a hallmark of various human diseases. Global hypomethylation can lead to genomic instability, loss of imprinting, and activation of repetitive or transposable elements, contributing to chromosomal rearrangements and oncogenic transformation [16]. Conversely, hypermethylation of promoter CpG islands is a frequent mechanism for the silencing of tumor suppressor genes in cancer, while altered gene body methylation has been implicated in inappropriate transcriptional activation of oncogenes [17]. Beyond cancer, dysregulated methylation landscapes have been associated with a wide spectrum of disorders, including neurodevelopmental and neurodegenerative diseases, autoimmune disorders, imprinting syndromes, and metabolic and cardiovascular pathologies [18], [19].

Because methylation integrates both genetic and environmental influences and is reversible, it serves as both a mediator of phenotypic plasticity and a valuable biomarker for diagnosis, prognosis, and therapeutic monitoring. Recent advances in long-read sequencing have further enabled

single-molecule, base-level methylation profiling in native DNA, opening new perspectives for integrative epigenomic analyses [20], [21].

1.1.3 Transcriptomics

Transcriptomics is the large-scale study of the transcriptome, defined as the complete set of RNA molecules expressed in a cell, tissue, or organism at a given time. Unlike the genome, which remains largely static, the transcriptome is highly dynamic, reflecting the real-time functional state of a biological system. By quantifying the abundance, diversity, and structure of RNA transcripts, transcriptomics provides critical insights into gene regulation, differentiation, and disease mechanisms.

Initially, transcriptome analysis relied on hybridization-based approaches such as microarrays, which provided global but probe-limited and semi-quantitative views of gene expression. The introduction of RNA sequencing (RNA-seq) revolutionized transcriptomics by enabling unbiased, genome-wide quantification of transcripts [22], [23]. RNA-seq workflows combine experimental and computational steps to map reads to reference genomes, quantify expression levels, and identify differentially expressed genes and transcript isoforms. This approach has become the gold standard due to its sensitivity, accuracy, and scalability [24].

Modern transcriptomics extends far beyond gene expression quantification. Alternative splicing can be systematically characterized, revealing how individual genes give rise to multiple isoforms with distinct functions. Strand-specific sequencing allows the detection of antisense transcription and non-coding RNAs, while ribosome profiling captures actively translated regions of the transcriptome, providing a direct readout of protein synthesis.

Epitranscriptomics

Similarly to DNA, RNA molecules can undergo various chemical modifications, collectively known as epitranscriptomic marks, including N6-methyladenosine (m6A), 5-methylcytosine (m5C), and pseudouridylation, among others. These modifications influence RNA stability, localization, splicing, and translation efficiency, thereby playing pivotal roles in regulating gene expression. However, despite their recognized importance, the systematic study of epitranscriptomics remains limited, largely due to technological constraints [25].

In this context, transcriptomics still represents a cornerstone of modern biology, enabling the exploration of gene regulation at multiple levels. At the same time, the emerging field of epitranscriptomics highlights how much remains to be discovered about RNA biology, and points to the need for novel experimental and computational strategies capable of capturing not only the sequence and abundance of transcripts but also their chemical modifications and functional states.

1.2 DNA sequencing evolution

The investigation of genomic and epigenomic variation has become a central challenge of modern biology and medicine. As discussed in previous sections, both genetic diversity (ranging from single-nucleotide polymorphisms to large-scale structural rearrangements) and epigenetic modifications (such as DNA methylation and histone remodeling) play pivotal roles in shaping cellular identity, developmental processes, and disease. Unraveling these multilayered mechanisms requires technologies capable of detecting not only the primary DNA sequence, but also subtle modifications and structural changes with high precision.

In this context, DNA sequencing represents the most powerful and versatile approach to interrogate biological systems at the molecular level. From its earliest implementations, it has provided unprecedented insight into genome organization and function, enabling the systematic characterization of genetic and epigenetic variability. Over the past five decades, sequencing technologies have evolved through successive generations, which dramatically improved accuracy, throughput, read length and cost-efficiency, transforming genomics from a specialized discipline into a cornerstone of modern biomedicine.

1.2.1 First-Generation Sequencing

The first generation of DNA sequencing technologies was introduced in the late 1970s with the development of two landmark methods: the Maxam-Gilbert chemical degradation approach and the Sanger chain-termination method [26], [27]. The latter, in particular, became the dominant technique and remained the standard for over two decades. Sanger sequencing is based on enzymatic chain termination during DNA synthesis: a DNA polymerase

extends a primer using a mixture of normal deoxynucleotides (dNTPs) and labeled dideoxynucleotides (ddNTPs), which lack the 3'-hydroxyl group required for elongation. Incorporation of a ddNTP irreversibly terminates the nascent strand, generating fragments of varying length that all share a common primer.

The automation of this process in the 1980s and 1990s through fluorescent labeling, laser detection, and multi-capillary electrophoresis greatly improved throughput and reproducibility, leading to the first large-scale genome sequencing projects, including the Human Genome Project [28].

Sanger sequencing remains one of the most accurate technologies available, with per-base error rates below 0.01% and read lengths of up to 1000 bp. However, it comes with several limitations involving throughput, scalability and cost. Each reaction produces only a single read, and even multi-capillary systems cannot achieve the throughput required for large-scale experimental designs such as whole-genome sequencing (WGS) or whole-exome sequencing (WES), which today represent the most common applications of DNA sequencing in research. Moreover, the cost per base is relatively high compared to more recent technologies, restricting the feasibility of genome-wide explorative analyses. For these reasons, Sanger sequencing usage is currently mainly limited to validation purposes, where the goal is reaching extremely high accuracy on one or few genomic positions.

1.2.2 Massively Parallel Sequencing

The advent of second-generation sequencing (also referred to as next-generation sequencing, NGS) in the mid-2000s represented a turning point in the history of genomics. These new platforms introduced the concept of massively parallel sequencing, allowing to generate millions to billions of reads in a single run and reducing the cost per base by several orders of magnitude. This shift transformed DNA sequencing from a specialized, resource-intensive technique into a routine tool capable of supporting large-scale genomic projects [29], [30].

NGS technologies all rely on a common principle: DNA is fragmented, ligated to platform-specific adapters, and clonally amplified on a solid support to generate spatially separated clusters of identical molecules. This amplification step ensures that the signal produced during sequencing is detected in parallel across the entire flow cell or plate.

Different platforms implement this principle with distinct chemistries. The Roche 454 system introduced pyrosequencing [31], in which nucleotide incorporation triggers a cascade of enzymatic reactions producing light; although capable of relatively long reads for its time, it suffers from miscalls in homopolymeric regions. Illumina’s sequencing-by-synthesis [32] rapidly became the dominant technology: fluorescently labeled reversible terminators are incorporated one base at a time, imaged, and chemically cleaved to allow the next cycle. This method enables highly accurate short reads, typically 100-300 bp, with low error rates and remarkable scalability. Lastly, the Applied Biosystems SOLiD platform adopted a sequencing-by-ligation approach, interrogating dinucleotide combinations through a series of ligation events, achieving very high accuracy at the cost of shorter reads and a complex workflow.

The advantages of NGS technologies are substantial. They provide massive throughput, generating gigabases to terabases of sequence data per run, enabling large-scale projects such as WGS, WES and RNA-seq, while also ensuring high base accuracy ($\sim 99.9\%$). Furthermore, the same platforms can be adapted to several applications, including whole-genome and whole-exome sequencing in large cohorts, transcriptome profiling through RNA-seq and genome-wide regulatory mapping (ChIP-seq, ATAC-seq) [33]. These approaches profoundly reshaped the study of gene regulation, genome variability, and disease mechanisms, laying the foundation for modern population genomics and precision medicine.

Despite these transformative advances, NGS platforms remain inherently constrained by their short read lengths. Reads of only few hundred base pairs struggle to resolve long repetitive regions, span structural variants, or reconstruct complex haplotypes, and amplification biases can introduce uneven coverage or allele dropout [34]. Moreover, because standard NGS workflows cannot directly detect DNA modifications or long-range chromatin features, additional experimental steps are required to study methylation or higher-order regulatory architecture [35].

Although part of these limitations have been overcome by the advent of long-read and single-molecule sequencing technologies (discussed in the following section), second-generation sequencing has remained the dominant platform especially for clinical applications because of its high accuracy, reproducibility and availability of standardized analysis workflows.

1.2.3 Long-Read Sequencing

Third-generation sequencing (TGS) emerged in the 2010s to address the key shortcomings of NGS: the inability to span complex genomic regions, assemble genomes *de novo* with high contiguity, or detect epigenetic modifications directly. In contrast to previous generations, TGS platforms sequence individual DNA or RNA molecules without prior amplification, in real time, and with read lengths extending from tens to hundreds of kilobases. This shift enables a comprehensive view of genomic architecture and long-range interactions. The two leading TGS platforms are PacBio's Single-Molecule Real-Time (SMRT) sequencing and ONT's nanopore sequencing.

SMRT sequencing is based on the use of zero-mode waveguides (ZMWs), nanostructures that confine the observation volume to thezeptoliter scale. Within each ZMW, a single DNA polymerase is immobilized at the bottom of the chamber and continuously supplied with fluorescently labeled nucleotides. Each of the four nucleotides carries a distinct dye attached to its terminal phosphate group. When the polymerase incorporates a nucleotide into the nascent DNA strand, the fluorophore remains transiently associated with the active site, generating a base-specific fluorescent pulse that is recorded in real time. Following incorporation, the labeled phosphate is cleaved and rapidly diffuses out of the detection zone, ensuring that background fluorescence does not accumulate and allowing the process to continue seamlessly. This continuous, real-time detection enables the generation of long reads, with average lengths exceeding 10 – 20 kb and maximum reads surpassing 100 kb [36].

SMRT sequencing can operate in two main modes [37], which reflect a trade-off between read length and accuracy. In the Continuous Long Read (CLR) mode, the polymerase extends along a linear template, producing uninterrupted reads that often exceed tens of kilobases. This configuration is particularly powerful for *de novo* genome assembly and structural variant discovery, as the extended read length allows coverage of repetitive regions and complex rearrangements. However, because each nucleotide is read only once, raw CLR data exhibit relatively high error rates, typically in the range of 10 – 15%, with errors dominated by false insertions and deletions. These errors can be mitigated by increasing sequencing depth and applying consensus correction across multiple reads, but they limit the direct application of CLR data in contexts where per-base accuracy is critical.

The alternative configuration, known as Circular Consensus Sequencing

(CCS) or High-Fidelity (HiFi), relies on the use of SMRTbell templates, in which double-stranded DNA fragments are ligated to hairpin adapters to create closed circular molecules [38]. When such a template is loaded into a ZMW, the polymerase can traverse the circular molecule repeatedly, generating multiple subreads of the same DNA fragment within a single sequencing run. These redundant readings are then computationally combined to produce a highly accurate consensus sequence, commonly referred to as a HiFi read. HiFi reads typically reach accuracies above 99% while still maintaining substantial lengths, often 10 – 15 kb, thereby combining the benefits of long-read sequencing with the accuracy traditionally associated with short-read technologies.

The choice between CLR and HiFi depends on the scientific question, although HiFi is largely more used. CLR data maximize read length and genome contiguity but require higher coverage and careful error correction, whereas CCS data offer superior base-level accuracy, making them highly suitable for applications such as variant detection, haplotype phasing, and clinical genomics. Together, these two operational modes illustrate the flexibility of SMRT sequencing in balancing throughput, read length, and accuracy to meet diverse research needs.

Nanopore sequencing represents a fundamentally distinct approach to nucleic acid analysis, relying on the direct measurement of biophysical signals rather than synthesis or labeling chemistries [39], [40]. In this method, individual DNA or RNA molecules are driven through nanoscopic pores embedded in an artificial membrane under the influence of an applied voltage, which serves a dual role: it establishes an ionic current across the membrane, which provides the measurable signal, and it simultaneously exerts an electrophoretic force on the negatively charged nucleic acid backbone, thereby inducing its translocation through the pore. As the nucleic acid traverses, the bases that occupy the pore at a given time partially obstruct the ionic current flowing through it. Each combination of nucleotides generates a characteristic disruption in current intensity, producing a complex signal that encodes the underlying sequence. These raw electrical traces are subsequently decoded by computational basecalling algorithms, which apply statistical modeling and, more recently, deep learning methods to translate the current patterns into nucleotide sequences with base-level resolution.

A defining feature of nanopore sequencing is its ability to read a nucleic acid molecule continuously from end to end. In practice, ONT platforms

often produce reads of hundreds of kilobases and ultra-long reads of over two megabases have been reported.

Additionally, ONT sequencing is very versatile thanks to the range of available devices, from the portable MinION, designed for field-based applications, to high-throughput instruments such as the PromethION, which can deliver hundreds of gigabases of sequence data in a single run.

The advantages of TGS are considerable. Long-read sequencing resolves repetitive regions, segmental duplications, and structural variants that are inaccessible to short-read technologies. It allows for accurate assembly of complex genomes *de novo*, haplotype phasing across long distances, and the comprehensive characterization of isoform diversity in transcriptomics. Furthermore, the ability to interrogate nucleic acids in their native state offers unique opportunities for integrative analyses of genome regulation, making TGS particularly powerful in functional epigenomics. In clinical contexts, the ability to detect structural rearrangements and long-range variants has opened avenues in oncology, rare disease diagnosis, and infectious disease surveillance [41].

However, TGS technologies also present several challenges. Early implementations of both SMRT and nanopore sequencing exhibited raw error rates in the range of 10 – 15%, substantially higher than those of NGS. Although the introduction of HiFi reads for PacBio and the rapid advances in ONT sequencing chemistry and basecalling have markedly improved accuracy, long-read data still display greater variability than Illumina sequencing, particularly in repetitive or homopolymeric regions. Throughput and cost per base have steadily improved but remain higher than those of second-generation platforms in many contexts. Nanopore sequencing further relies on complex signal processing and machine learning-based basecalling, whose performance depends strongly on the chemistry, pore design, and flow cell quality, leading to variability in accuracy and yield across experiments. PacBio SMRT sequencing, while capable of generating highly accurate HiFi reads, is constrained by reduced read length in HiFi mode and requires substantial investment in specialized instrumentation. Despite the high base-level accuracy achieved by PacBio HiFi reads, important limitations remain for epigenomic and epitranscriptomic applications. In particular, although recent methods for single-molecule methylation detection from CCS data represent significant progress, reported accuracies remain below 85% [42], [43]. Moreover, with the exception of a few specialized protocols [44], direct

RNA sequencing is currently feasible only with nanopore technology.

In summary, third-generation sequencing represents a paradigm shift in genomics, enabling single-molecule, long-read, and epigenetically informative sequencing. While technical challenges remain, continuous improvements in chemistry, instrumentation, and computational methods have made TGS indispensable for applications where long-range genomic information is critical. Together with the scalability of second-generation technologies, TGS platforms contribute to an increasingly diverse sequencing landscape that is driving forward both fundamental research and clinical translation in the omics era.

Chapter 2

Aim of the Thesis

TGS technologies have reshaped genomics by enabling the direct sequencing of individual DNA or RNA molecules without PCR amplification. Platforms such as ONT and PacBio generate long reads that can span entire genes, transcripts, and structural variants, allowing the reconstruction of genomic regions that are difficult or impossible to resolve with short-read approaches. Beyond their extended read length, these technologies also enable the direct detection of nucleotide modifications, integrating structural, epigenetic, and transcriptomic information within a single assay. As a result, long-read sequencing provides a unified framework for multiple layers of genome regulation: from the accurate characterization of large structural variants, to the profiling of DNA methylation without chemical conversion, and the sequencing of native RNA molecules while preserving their modifications.

Despite the rapid technological progress and the unique advantages offered by long-read sequencing, several technical and practical challenges still limit its widespread adoption and have so far prevented it from replacing NGS as the standard technology in most genomic applications. In clinical and diagnostic settings, short-read sequencing remains the preferred method due to its robustness, standardization and regulatory maturity.

From a technological standpoint, long-read platforms still exhibit higher error rates than short-read technologies, although accuracy has improved substantially in recent years, particularly for ONT-based approaches. This intrinsic noise complicates the detection of small variants and the quantitative analysis of DNA or RNA base modifications, especially in repetitive

regions or in loci with limited coverage. Throughput and cost also remain important considerations. PacBio sequencing is fundamentally constrained by the fixed number of ZMWs available on each SMRT Cell, which limits the total number of simultaneously observable polymerase complexes and therefore the achievable yield. In contrast, ONT devices offer higher potential throughput but suffer from substantial run-to-run variability, as the number of active pores and their stability over time can vary widely across experiments. These factors make sequencing output less predictable and can complicate study design, especially for large cohorts.

From a computational perspective, the field is evolving rapidly, with an increasing demand for robust, efficient, and interpretable algorithms designed specifically for long-read data. Many existing tools originated as adaptations of short-read methods, whose underlying assumptions do not always align with the error profiles, read lengths, or signal characteristics of nanopore and SMRT sequencing. Developing analytical frameworks that fully exploit the unique properties of long-read technologies is therefore a key step toward enabling their reliable use in research, diagnostics, and precision medicine.

Within this framework, the research presented in this thesis focuses on the design and implementation of computational methods aimed at enhancing the analysis and interpretation of long-read sequencing data across multiple genomic and epigenomic applications.

The work conducted for this thesis is structured around three main research directions, each addressing a distinct analytical challenge and leveraging the strengths of long-read sequencing:

1. **Structural variant detection.** I contributed to the development of a novel computational method for the detection of structural variants from nanopore sequencing data. The proposed approach integrates read-level evidence with local assembly-based refinement, significantly improving accuracy in breakpoint localization and reducing the rate of false positives compared to widely used SV callers. This work demonstrates how long-read information can be used not only to detect complex rearrangements with higher confidence, but also to reconstruct the precise architecture of variant haplotypes.
2. **Differential DNA methylation detection.** I developed an innovative computational framework for the identification of differentially methylated regions (DMRs) from nanopore sequencing data. The

method, implemented in the PoreMeth2 package, introduces a bivariate segmentation model that simultaneously analyzes changes in methylation frequency and in epiallelic entropy, enabling the distinction between DMRs generated by stochastic epigenetic drift and those arising from clonal epiallelic selection. This approach provides a more comprehensive view of methylation dynamics, uncovering both the directionality and the internal variability of methylation changes, and offering new insights into the functional consequences of epigenetic remodeling.

3. Direct RNA sequencing and detection of RNA modifications.

Finally, I conducted a preliminary investigation into the signal processing challenges associated with Nanopore Direct RNA sequencing. Rather than developing a finalized detection workflow, this section of the thesis presents an exploratory analysis focused on the stability of raw signal extraction. By comparing basecaller-guided segmentation against explicit signal re-alignment, I assessed the impact of different alignment strategies on signal reproducibility. The observations suggest that while re-alignment appears necessary to mitigate model-dependent variability, read-guided alignment warrants further investigation as a potential strategy for characterizing k-mer variance. These initial findings offer a partial basis for future research into unsupervised, control-free frameworks for epitranscriptomic analysis.

Collectively, this thesis aims to advance the computational maturity of long-read sequencing technologies. By introducing novel algorithmic frameworks for structural and epigenetic profiling, and by assessing emerging capabilities in direct RNA analysis, this work strives to translate raw technological potential into actionable biological insights. The overarching objective is to facilitate the transition of long-read sequencing from a specialized research tool to a cornerstone of routine genomic and diagnostic medicine.

Chapter 3

Structural Variant detection

*SV discovery represents one of the primary applications of long-read sequencing, where the extended read length offers unique advantages for resolving large and complex genomic rearrangements. Despite significant progress, the accurate and comprehensive detection of SVs from nanopore data remains a challenging task, due to high sequencing error rates, signal noise, and the difficulty of distinguishing true events from alignment artifacts. In this chapter, we present GASOLINE, a novel computational method for SV detection from Oxford Nanopore sequencing data. GASOLINE introduces an alternative clustering strategy designed to improve breakpoint precision and reduce redundant calls, while also providing a dedicated framework for identifying somatic structural variants through comparative analysis of matched sample pairs.*¹

¹This chapter has been published as "GASOLINE: detecting germline and somatic structural variants from long-reads data" in Scientific Reports, 2023 [45]

3.1 Background

Structural Variants (SVs) are genomic alterations that involve segments of DNA typically larger than 50 base pairs. They represent one of the most extensive and functionally significant forms of genomic variation, encompassing diverse rearrangement types such as:

- **Copy Number Alterations (CNAs)**, referring to the gain (duplication) or loss (deletion) of one or more copies of a genomic segment or, in extreme cases, entire chromosomes (aneuploidies);
- **Insertions**, which introduce novel DNA sequences not present in the reference genome, often originating from mobile elements, tandem repeats, or exogenous integrations;
- **Inversions**, where a DNA segment is excised and reinserted in reverse orientation, potentially disrupting coding regions or altering the three-dimensional chromatin context of regulatory elements;
- **Translocations**, involving the exchange of genomic material between non-homologous chromosomes, which may be either balanced (without net gain or loss of DNA) or unbalanced (leading to copy number changes);
- **Complex rearrangements**, comprising multiple concurrent events (e.g., chromothripsis, chromoplexy, templated insertions) that result in highly rearranged and mosaic genome architectures.

Although SVs are a natural component of inter-individual genomic diversity and contribute to adaptive evolution, their occurrence can have profound pathological consequences. Depending on their type, size, and genomic location, SVs may disrupt coding sequences, alter gene dosage, or perturb long-range regulatory interactions. They have been implicated in a broad spectrum of human diseases, including congenital malformations, neurodevelopmental and neuropsychiatric disorders, and various cancers [46]–[49].

3.1.1 Cytogenetics and Hybridization-based techniques

The study of SVs has evolved alongside technological advances in genomics. The earliest insights into large-scale chromosomal rearrangements arose from classical cytogenetics: conventional karyotyping, based on Giemsa-stained metaphase chromosomes, enabled the visualization of aneuploidies, translocations, inversions, and large deletions or duplications. Despite being a cornerstone of cytogenetics, this approach offers limited resolution (several

megabases), requires cell culture, and fails to detect sub-microscopic rearrangements [50].

The advent of molecular cytogenetics provided more refined methods, based on sequence hybridization, that improved the detection of CNVs down to a resolution of tens of kilobases. Specifically, fluorescence in situ hybridization (FISH) [51] partially overcame the limitations of standard karyotype by employing fluorescently labeled DNA probes to complementary genomic regions. FISH markedly improved sensitivity but remained locus-specific, requiring prior knowledge of the target sequence and thus unsuitable for genome-wide screening.

The development of microarray-based techniques, particularly chromosomal microarray analysis (CMA) and array comparative genomic hybridization (aCGH) [52], represented a substantial advancement. By hybridizing differentially labeled test and reference DNA to genomic probe arrays, these methods could detect CNVs at increasingly fine resolutions (down to tens of kilobases in optimal conditions). However, they remained blind to balanced SVs, such as inversions and translocations, and had limited breakpoint precision.

3.1.2 NGS-based SV detection

The advent of NGS technologies marked a fundamental shift in SV discovery, enabling genome-wide detection with nucleotide-level resolution. Unlike previous cytogenetic methods, NGS-based SV detection relies on three primary computational approaches:

- **Paired-end mapping:** This strategy infers structural changes from deviations in the expected alignment of forward and reverse read pairs, specifically analyzing discrepancies in their mapping distance (insert size) or relative orientation. It is versatile and can potentially detect any type of SV.
- **Split-read mapping:** This approach aligns separate portions of a single read to distinct genomic loci, allowing for the identification of breakpoints with base-pair precision, which is particularly effective for smaller events.
- **Read-depth analysis:** Also known as coverage-based methods, these algorithms quantify local fluctuations in sequencing depth to identify Copy Number Alterations (CNAs), such as deletions and duplications.

While integrating these signals significantly expanded the dynamic range of SV detection [53], the use of short-read sequencing imposes intrinsic limitations on SV discovery. Primarily, the reconstruction of structural rearrangements is an indirect process inferred from mapping signatures rather than directly observed in the data. Insertions larger than the read or fragment length (typically 300-600 bp for paired-end sequencing) cannot be fully captured, since they do not appear as continuous inserts within individual reads. As a consequence, large novel insertions or mobile element insertions are often invisible to short-read approaches, unless represented in the reference genome or captured by assembly-based strategies.

Similarly, split-read mapping, which is essential for resolving breakpoints at base-pair resolution, becomes unreliable in repetitive or low-mappability regions. When reads are only 100-150 bases long, even small repeats or paralogous sequences can generate multiple equally valid alignments, preventing the unique placement of split fragments and leading to ambiguous breakpoint predictions. These limitations are exacerbated in regions rich in segmental duplications or tandem repeats, where short reads fail to span the full extent of the rearrangement.

As a result, short-read based methods often produce fragmented or incomplete reconstructions of complex or nested structural variants, and are biased toward the detection of deletions and small duplications. The inability to directly observe rearrangements within single reads fundamentally constrains the resolution and completeness of SV detection achievable with short-read sequencing technologies. While these downsides can be partially mitigated by assembly-based strategies, generating contiguous and correctly phased haplotypes from short reads remains technically challenging, particularly in structurally complex regions, where assembly errors and gaps are frequent.

3.1.3 TGS-based SV detection

The introduction of long-read sequencing, first through PacBio’s SMRT technology and later through ONT, profoundly changed this landscape. By generating reads spanning up to hundreds of thousands of base pairs, these platforms enable direct observation of structural rearrangements, often within single reads.

Studies leveraging PacBio data demonstrated that up to 80% of SVs identified by long reads were missed by short-read approaches [2], [54], [55],

underscoring the importance of read length for comprehensive variant discovery. Long reads drastically enhance breakpoint precision, facilitate phasing of compound variants, and allow characterization of repetitive or highly polymorphic genomic regions previously inaccessible to NGS.

Despite these strengths, challenges remain in terms of per-base accuracy, computational costs, and the design of specialized algorithms able to exploit the full potential of long-read data. The next section provides an overview of existing computational methods for structural variant detection from nanopore sequencing and introduces GASOLINE, a novel algorithm designed to improve accuracy and reduce false positives in SV calling.

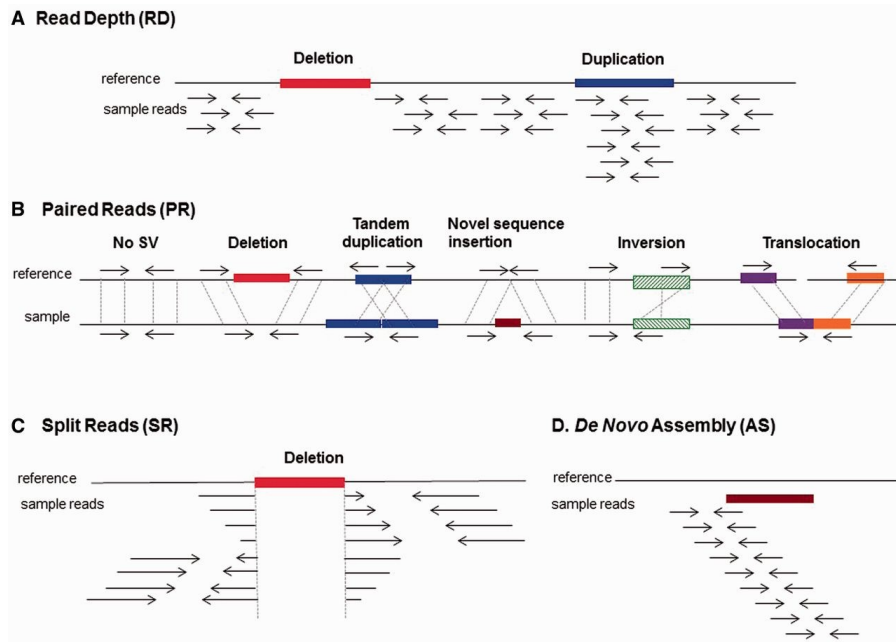


Figure 3.1: The figure represents standard approaches for SV detection from NGS data, mainly categorized in assembly-based (D) and alignment-based (A, B, C), which can be further subcategorized in read-depth based (A), read-pair based (B) and split-read based (C). [50]

3.2 State of the art

The advent of long-read sequencing technologies has profoundly transformed SV discovery, enabling direct observation of large-scale genomic rearrangements with a resolution that was previously unattainable with short-read data. Over the past few years, a growing number of computational tools have been developed to exploit the unique properties of long-read sequencing and capacity to span repetitive elements and complex breakpoints within individual reads. Despite sharing the same goal, these methods differ substantially in their analytical strategies, the types of SVs they target, and their computational performance [56].

Broadly, SV detection algorithms for long-read data can be grouped into two main paradigms: alignment-based and assembly-based approaches (Figure 3.2, [57]).

Alignment-based methods rely on mapping individual reads to a reference genome to identify characteristic mapping signatures indicative of structural rearrangements. These include split alignments (where a read aligns to disjoint genomic loci), supplementary alignments, and significant soft-clipped bases or intra-read gaps relative to the reference. Once such signatures are collected, reads supporting the same event are clustered, and a consensus SV call is generated. This approach is computationally efficient and scalable; however, its sensitivity relies heavily on the accuracy of the aligner and may decline in highly repetitive regions or when resolving large insertions that exceed the read length or are entirely absent from the reference.

Assembly-based methods, in contrast, reconstruct contiguous genomic sequences (or contigs) from long reads, either through *de novo* or reference-guided assembly, and subsequently align them to the reference to identify structural divergences. This strategy is powerful for detecting large novel insertions and resolving complex, nested rearrangements that are difficult to infer from read alignments alone. However, it entails significant trade-offs. Assembly is computationally intensive and typically requires high sequencing coverage to be reliable. Furthermore, while modern "diploid-aware" assemblers strive to separate haplotypes, correctly phasing structural variants across long genomic distances remains challenging without auxiliary data (e.g., parental genomes or Hi-C), often leading to haplotype switching errors or chimeric contigs. Crucially, assembly approaches generally aim to reconstruct the dominant sequences in a sample, making them suboptimal

for detecting low-frequency somatic variants in heterogeneous samples, such as tumor tissues. In these contexts, identifying subclonal rearrangements requires the sensitivity of read-level alignment approaches, which preserve the information contained in individual molecules representing rare cell populations.

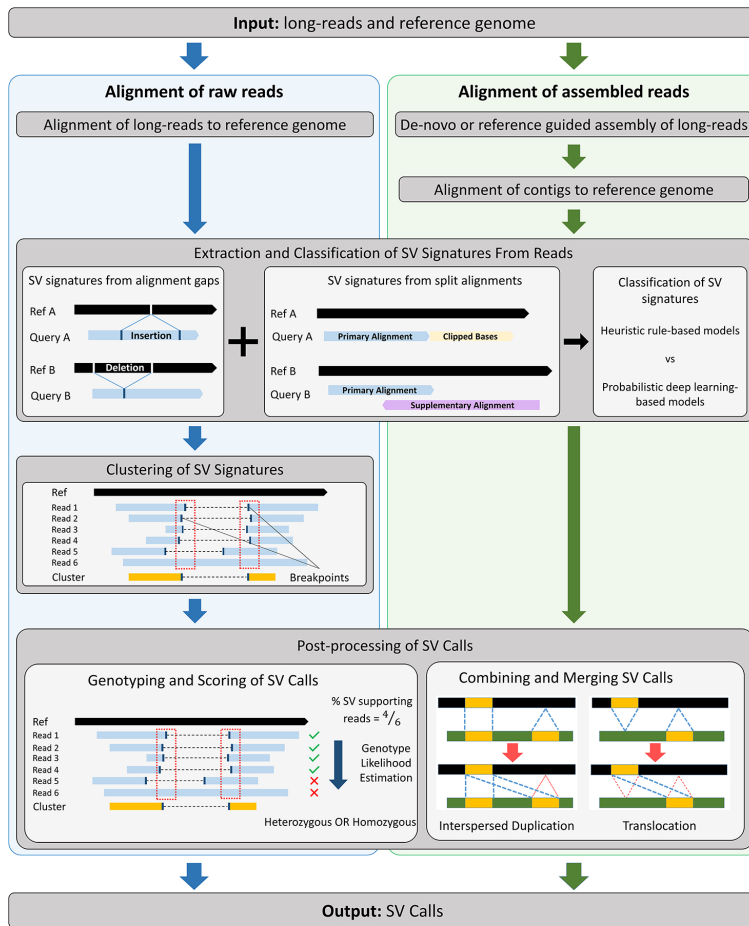


Figure 3.2: Summary of the workflow of standard approaches for SV detection from TGS aligned reads (left) and assembled contigs (right) [57].

To balance sensitivity and accuracy, some modern SV callers adopt a hybrid strategy, combining alignment-derived signatures for initial detection

with local reassembly or consensus refinement around candidate breakpoints. This paradigm is implemented in several state-of-the-art tools for ONT and PacBio data: Sniffles2 (which is also included in automated pipelines for both technologies) [58], cuteSV [59] and SVIM [60] follow this paradigm and are widely considered as standard methods for SV detection in ONT and PacBio datasets.

However, these tools still face challenges in achieving consistent accuracy across variant classes and sequencing conditions. Benchmark studies have shown that even the best-performing algorithms can differ markedly in their detection sensitivity, particularly for insertions, complex rearrangements, or events in repetitive regions, and that their performance is often highly dependent on the aligner used to map reads [61]. Moreover, breakpoint precision remains an active area of improvement as the high error rates cause signature coordinates from the same SV to be imprecise, often spanning even tens of base pairs [62], [63].

Importantly, only a very limited number of existing tools provide dedicated functionality for the detection of somatic structural variants. Among those that do, most are not among the best-performing or most widely adopted SV callers, often lacking the accuracy or robustness required for reliable use in heterogeneous or cancer-derived samples. For instance, while Sniffles2 implements a somatic mode, this function does not rely on direct comparison between matched sample pairs (e.g., tumor versus normal), but instead infers putative somatic variants from deviations in allelic fraction within a single sample. As a result, it provides only an approximate estimate of somatic events rather than a true pairwise comparison. The absence of accurate and efficient algorithms specifically designed for long-read based somatic SV discovery thus remains a major limitation in current computational genomics pipelines.

Within this evolving landscape, we developed GASOLINE to address some of these outstanding challenges and specifically, to enhance the resolution and reliability of SV detection while preserving information on allelic composition. By combining robust clustering of split-read signatures with localized sequence reconstruction, GASOLINE bridges the gap between raw read evidence and the accurate structural representation of genomic rearrangements, ultimately extending the interpretive power of long-read sequencing in both research and clinical genomics.

3.3 GASOLINE

Despite the remarkable progress in SV discovery enabled by long-read sequencing technologies, the accurate identification and characterization of genomic rearrangements remain challenging. Current algorithms typically rely on interpreting a combination of read alignment signatures to infer the presence of structural changes and to delimit their breakpoints. While these methods have greatly expanded our ability to interrogate genomic architecture, their performance is still limited by several intrinsic factors, including alignment errors, coverage heterogeneity, and the presence of repetitive or low-mappability regions. These sources of uncertainty often result in redundant or fragmented calls, low breakpoint precision, and poor agreement among tools. Within this context, we developed GASOLINE (Genomic structural variant Assembly, Sensing and OverLap-based INferencE), a computational framework designed to enhance the detection, clustering, and classification of structural variants from ONT sequencing data.

Similarly to other alignment-based strategies, the rationale underlying GASOLINE is that each long read aligned to a reference genome encodes local signals of potential rearrangements, which can be extracted and systematically integrated across multiple reads. These signals are represented as candidate breakpoints, derived from abrupt changes in alignment coordinates, strand orientation, or mapping quality. Unlike conventional clustering strategies that rely on fixed genomic distance thresholds to group these breakpoints, GASOLINE introduces a similarity metric called Normalized Reciprocal Overlap (NRO), which quantifies the degree of overlap between two genomic intervals in a scale-independent manner. This metric allows for an adaptive grouping of breakpoints even in the presence of mapping imprecision or read-to-read variability, a common issue in nanopore data. The NRO between two intervals i and j is defined as:

$$NRO_{ij} = 2 \cdot \frac{Overlap}{Li + Lj} - \frac{NO_i + NO_j}{NO_{Norm}}$$

where:

- Li and Lj denote the lengths of the two intervals;
- $Overlap$ the size of their intersection;
- NO_i and NO_j the size of the non-overlapping regions of i and j .

The first term corresponds to the classical reciprocal overlap, capturing the proportion of shared bases between the two intervals. The second term is a correction factor that penalizes large discrepancies in size between intervals, normalized by a constant parameter $NOnorm$, empirically determined to balance the two contributions. This correction prevents the erroneous clustering of regions that exhibit minimal overlap but substantial length differences, a common issue in conventional reciprocal overlap schemes.

By combining both spatial proximity and size similarity, the NRO yields a normalized score ranging from 0 (no correspondence) to 1 (perfect match), enabling GASOLINE to achieve highly accurate and robust clustering of breakpoint signals across reads. In contrast to widely used tools such as Sniffles2, cuteSV, and SVIM, which employ standard distance-based rules or heuristic clustering windows, the NRO metric inherently accounts for local variability in breakpoint precision and alignment uncertainty. This design improves the detection of variants in repetitive regions or areas of complex genomic organization, where small shifts in alignment can otherwise lead to multiple spurious calls representing the same event. During the clustering phase, GASOLINE merges intervals with an NRO above a user-defined threshold (Thr, ideally between 0.6 and 0.7), ensuring that only spatially coherent and well-supported signals are retained. This adaptive measure replaces rigid coordinate-based distance rules, improving the sensitivity and robustness of clustering across genomic contexts.

3.3.1 Workflow

The general workflow of GASOLINE proceeds through three main stages (Figure 3.3).

1. In the first stage, reads are scanned to identify alignment patterns indicative of structural variation: SV signatures (gaps, soft-clipped segments and orientation switches) are extracted from the CIGAR string in alignment files and interpreted as potential breakpoints. Each detected signal is encoded in a standardized representation that retains its genomic coordinates, event type, alignment score, and supporting read. For each overlapping pair of signatures ij , NRO_{ij} is calculated and all NRO values are then used to build an adjacency matrix for the whole signature group.

2. The second step consist in NRO-based clustering, in which the adjacency matrix is used to build an undirected graph where each signature is a node and edges are NRO values between node pairs, and only exist if $NRO \geq Thr$. In the clustering step, GASOLINE employs the Eppstein-Löffler-Strash algorithm [64] to select a subset of reads such that every node is linked with all the others (fully connected nodes).
3. Lastly, each candidate variant is genotyped by means of a maximum-likelihood Bayesian classification algorithm based on the fraction of reads showing variant signatures [65].

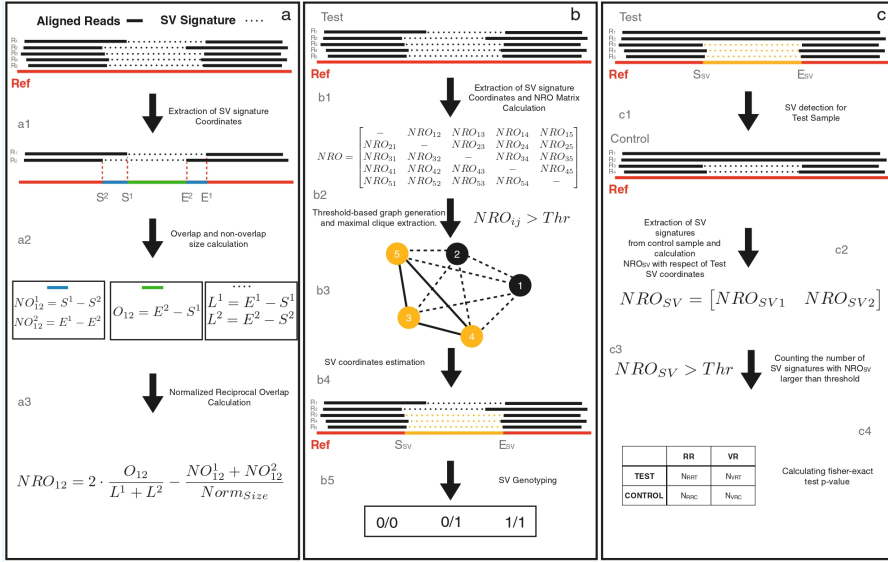


Figure 3.3: The figure represents the workflow behind GASOLINE germline (panels a, b) and somatic (panel c) SV calling.

3.3.2 Somatic SV calling

Another relevant advancement of GASOLINE with respect to other methods, is the possibility to identify somatic SVs from sample pairs. This is a particularly challenging task especially in cancer genomes, as variants are often subclonal with variable allelic fractions, and embedded in highly rearranged genomic contexts. Traditional germline-oriented SV callers are not

optimized for such conditions, as they generally assume uniform zygosity and lack dedicated modules for paired-sample analysis.

The somatic module of GASOLINE addresses these limitations by leveraging the NRO metric to compare SV signatures detected in tumor and matched-normal samples. Somatic SV discovery is performed by first applying the germline detection module to the tumor dataset (test), generating a comprehensive list of candidate SV signatures. These clusters are then cross-referenced with signatures extracted from the normal sample (control), and their degree of overlap is quantified through the NRO: candidate events exhibiting an $NRO \geq Thr$ are classified as shared between the two datasets and are therefore filtered out, while those below the threshold are retained as tumor-specific candidates.

Finally, the statistical significance of each putative somatic SV is evaluated by comparing the relative abundance of supporting and reference reads in the tumor and control samples through a Fisher’s exact test. Variants with a p -value below the significance threshold are classified as bona fide somatic events. This statistical filtering step allows the removal of spurious candidates arising from sequencing noise or stochastic fluctuations in read coverage, ensuring that only confidently tumor-specific rearrangements are retained and therefore reducing false positive calls.

3.4 Results

3.4.1 Germline variants

To evaluate the accuracy and robustness of GASOLINE in the detection and genotyping of germline SVs, we performed several tests on both synthetic and real long-read datasets. These experiments were designed to assess the method’s ability to detect all SV subtypes (deletions, insertions, inversions, duplications, and translocations) across a wide range of sequencing coverages and error conditions.

We first generated synthetic datasets using PBSIM2 [66], simulating long-read data with average read length of ~ 10 kb and a global accuracy of 90%, therefore resembling the characteristics of both PacBio and ONT reads. Reads were generated to produce different datasets with coverage ranging from 5x to 30x. Deletions, insertions, duplications, inversions, and translocations were simulated in both homozygous and heterozygous states, with

sizes ranging from ~ 50 bp to ~ 5 kb by producing simulated reads using a modified version of the hg19 reference genome (see Methods). Synthetic reads were then aligned to the reference genome using two different read aligners (minimap2 [67] and NGML [68]) to also evaluate the robustness of GASOLINE across different mapping strategies.

To investigate the influence of the NRO threshold on clustering performance, GASOLINE was tested using different NRO cutoffs (0.5, 0.6, 0.7, 0.8, and 0.9), and precision and recall were computed following standard definitions (see Figure 3.4). These analyses demonstrated that varying the NRO threshold has only a marginal effect on overall performance. However, optimal values depended on the alignment algorithm: NGMLR alignments performed best with lower thresholds (0.5-0.7), while minimap2 alignments achieved the best scores with higher thresholds (0.8-0.9), particularly for deletions. These findings demonstrated that GASOLINE's NRO-based clustering procedure maintains stable accuracy across a broad range of parameter settings.

We benchmarked GASOLINE (with a NRO threshold 0.8) against the three aforementioned state-of-the-art SV callers, Sniffles2, cuteSV, and SVIM. To this aim we employed real ONT and PacBio sequencing data of the well-characterized NA24385 individual (Ashkenazi trio) released by the Genome in a Bottle (GIAB) Consortium, for which a gold-standard set of high-confidence germline insertions and deletions has been released [69]. However, since other SV classes are not represented in this set, we complemented it with synthetic inversions, duplications, and translocations simulated with SURVIVOR [70] (see Methods). Both ONT and PacBio reads were down-sampled to simulate coverage levels from 5x to 30x and aligned using both minimap2 and NGMLR. We then compared GASOLINE's performance with that of Sniffles2, cuteSV, and SVIM by computing precision, recall, and F1-score as a function of the number of supporting reads per variant and of SV size.

Concerning simulated variants, across both ONT and PacBio data, GASOLINE and Sniffles2 consistently achieved the highest F1-scores, especially at lower sequencing coverages (5-10x), indicating that its NRO-based clustering can effectively group noisy and imprecise signatures derived from long-read alignments (Figure 3.5, panels a-f). Increasing coverage beyond 15x resulted in only marginal improvements, in agreement with previous reports that long-read SV detection saturates in performance above this level [61].

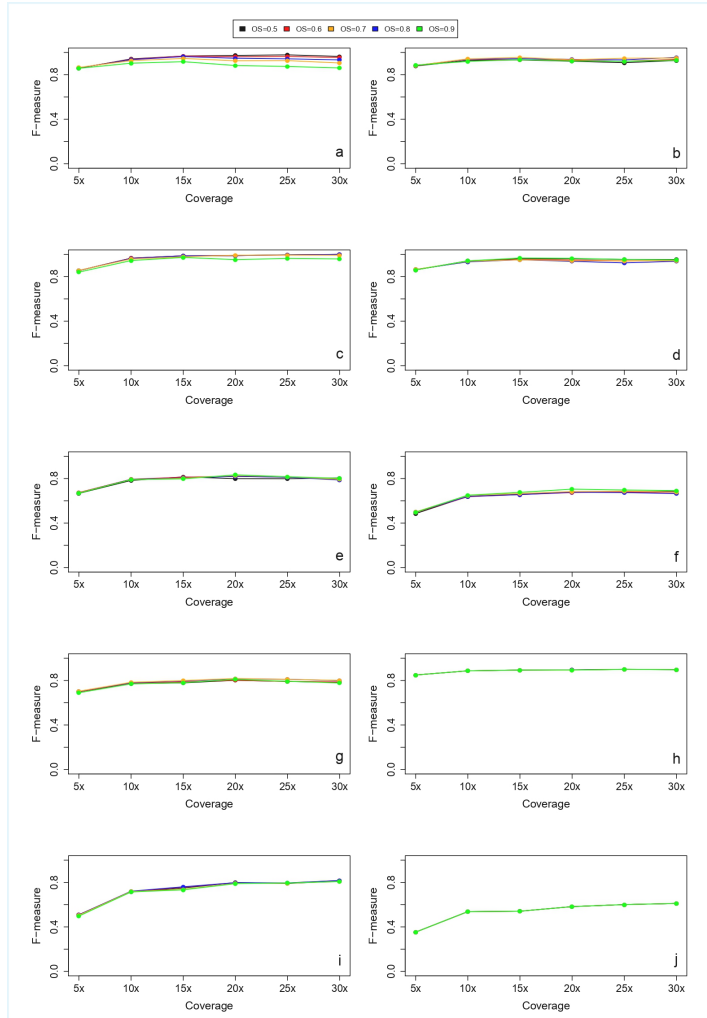


Figure 3.4: The figure shows the performance of GASOLINE with different NRO threshold values on synthetic data as a function of sequencing coverage. F1-scores are shown for different SV classes (deletions, panels a-b; insertions, panels c-d; duplications, panels e-f; translocations, panels g-h), different sequencing coverages (5x, 10x, 15x, 20x, 25x, 30x) and different read aligners (minimap2, panels a, c, e, g, i; NGMLR panels b, d, f, h, j). GASOLINE was tested for different NRO values thresholds (0.5, 0.6, 0.7, 0.8, 0.9, see color legend).

For large insertions and deletions (> 500 bp), all methods showed comparable results. However, for small SVs (< 500 bp), GASOLINE significantly outperformed all competitors, achieving the highest precision at comparable recall across all coverages, sequencing technologies, and aligners. This result highlights the effectiveness of the NRO-based clustering in discriminating true signatures from alignment noise in all ranges of SV sizes, while standard RO is less effective for small-sized variants (Figure 3.5, panels g-j).

As a last evaluation, we examined the genotyping performance of GASOLINE by stratifying SVs into heterozygous and homozygous classes. As observed for variant detection, for large variants all tools performed similarly, while for small SVs, GASOLINE consistently achieved the highest F1-measures across all datasets and coverage levels (Figure 3.5, panels k-n).

3.4.2 Somatic variants

To evaluate the ability of GASOLINE to identify somatic SVs, we first tested the algorithm on a real cancer dataset with matched normal control. For this purpose, we used the long-read sequencing data of the COLO829 melanoma cell line and its corresponding normal counterpart COLO829BL, which represent one of the few publicly available benchmarks for somatic SV detection. These datasets have been extensively characterized by multiple technologies, including Illumina short reads, PacBio and ONT long reads, and Bionano optical mapping, thus providing a reliable reference set of somatic rearrangements [71].

Both COLO829 and COLO829BL ONT and PacBio reads, obtained from EGA (see Methods), were aligned to the hg19 version of the human reference genome using both minimap2 and NGMLR achieving a mean coverage of approximately 60x for ONT and 50x for PacBio. We then performed somatic variant calling with GASOLINE and compared the results with those of the three tools used by Valle-Inclan et al. We observed that, by only retaining statistically significant variants (with different p thresholds), we strongly reduced the amount of false positive calls and increased precision (in the range of 0.6-0.8 for different SV types) while only slightly reducing recall, obtaining an overall better performance compared to the other state-of-the-art methods (Figure 3.6, panels e-o).

More importantly, with a threshold of $p < 0.001$ GASOLINE identified

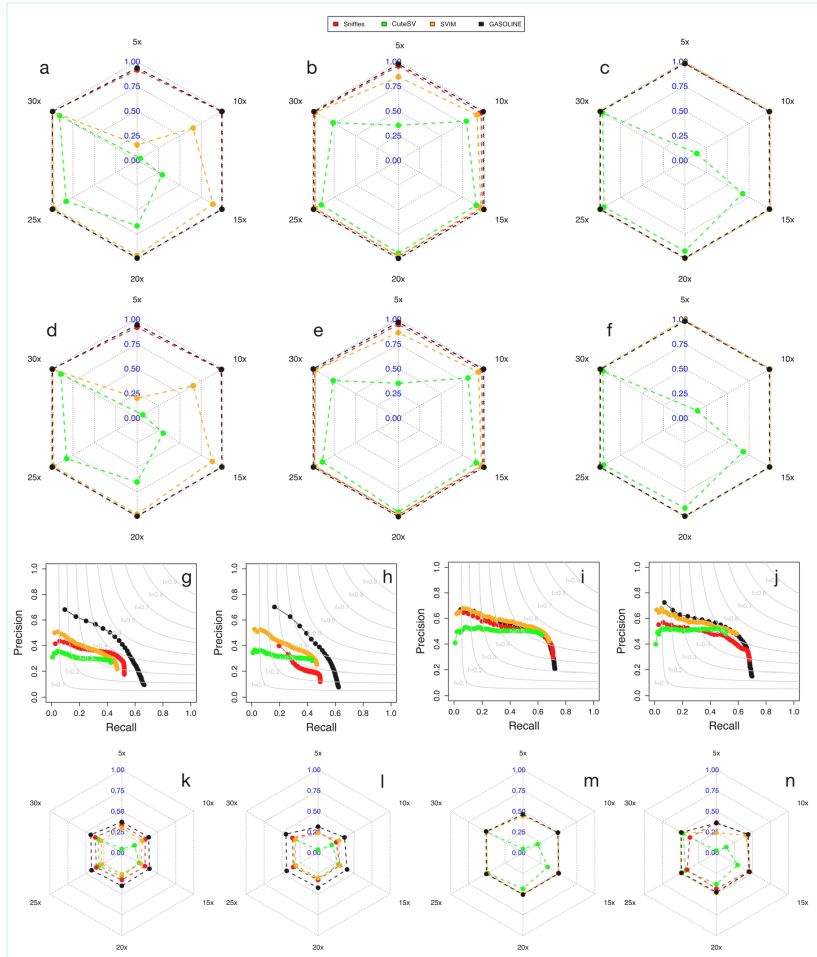


Figure 3.5: Performance of GASOLINE benchmarked against the other three tools for the detection of synthetic and real germline SVs. Panels a-f report F1-scores obtained for the identification of simulated inversions (panels a, d), duplications (panels b, e) and translocations (panels c, f) for both ONT (panels a-c) and PacBio (panels d-e) synthetic reads aligned with minimap2. In panels (g-j) we report precision and recall scores for each tool in the detection of insertions and deletions from the NA24385 dataset. Results are reported separately for small (< 500 bp; panels g, h) and large SVs (> 500 bp; panels i, j), from both ONT (panels g, i) and PacBio (panels h, j) data as a function of number of supporting reads, with precision and recall calculated including SVs with decreasing number of supporting reads. Panels k-n display F1-scores obtained by the four tools for the genotype assignment of small (panels k-l) and large SVs (panels m-n) with ONT (panels k, m) and PacBio (panels l, n) datasets at different sequencing coverages.

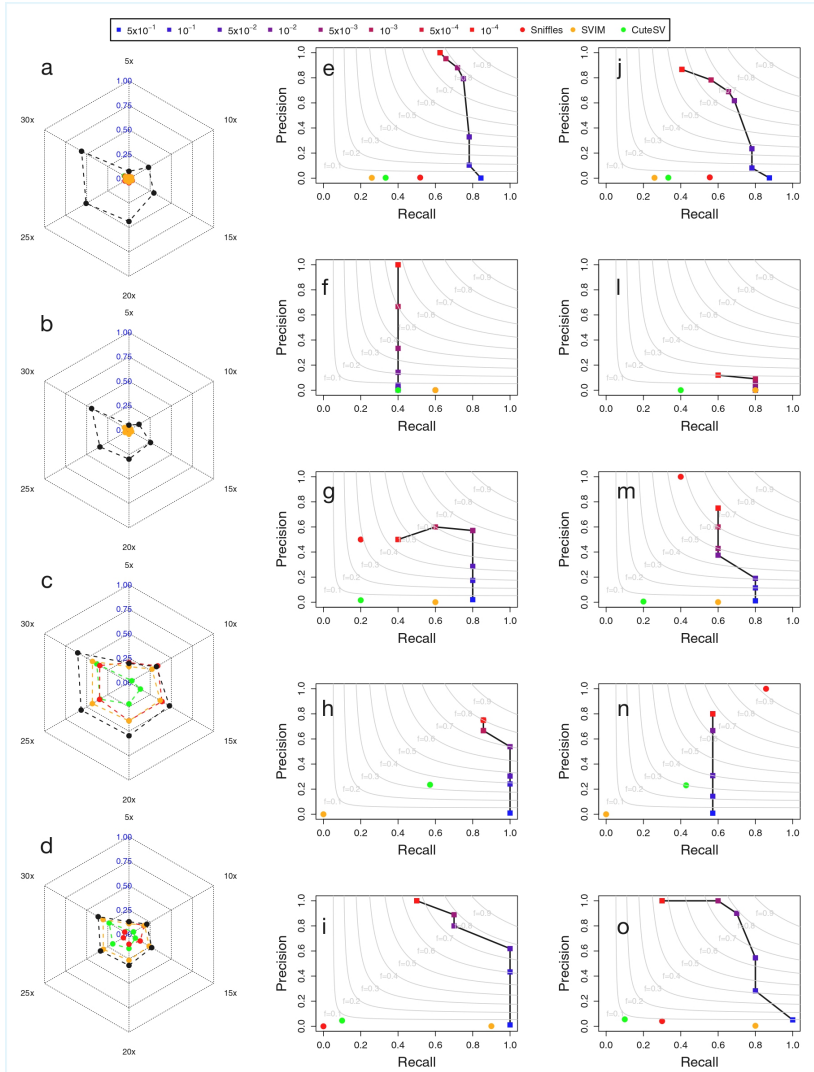


Figure 3.6: The figure shows the performance of GASOLINE in the identification of synthetic and real somatic variants. Panels a-d display F1-scores obtained by GASOLINE, Sniffles2, cuteSV and SVIM for the detection of small (< 500 bp, panels a-b) and large (> 500 bp, panels c-d) insertions and deletions from ONT (panels a, c) and PacBio (panels b, d) data. Panels e-o show F1-scores obtained by the same tools in the detection of different SV subtypes (deletions, panels e, j; insertions, panels f, l; duplications, panels g, m; inversions, panels h, n; translocations, panels i, o) from ONT (panels e-i) and PacBio (panels j-o) data as a function of p-value threshold used to filter out non significant variants (see color legend).

49 SVs of different types, 6 of which (1 deletion, 1 insertions, 1 duplication and 3 inversions) were classified as false positive as they were not reported by any of the other tools. However, we confirmed the somatic signatures with visual inspection of aligned reads with Samplot [72], and the same signatures were also supported by PacBio and NGS data for the same samples (Figure 3.7).

These results demonstrate the capability of GASOLINE to detect somatic SVs with high sensitivity and precision, even in complex genomic contexts. Nonetheless, despite the robustness of the COLO829 benchmark, it is important to note that this dataset still presents some limitations mostly ascribed to low representation of some SV types (e.g. only three insertions are reported in the truth set) and sizes, as the majority of variants are over 50 kb. For these reasons, we complemented the real data analysis with a controlled evaluation based on simulated datasets.

To this end, we selected 330 variants (insertions and deletions of sizes in the ranging from 50 bp to 5 kb) from the truth set for sample NA24385 and splitted ONT and PacBio reads with and without variant signatures in two separate alignment files to simulate somatic and germline samples respectively. We obtained sequencing data of $\sim 30x$ for both ONT and PacBio, which we aligned to the hg19 reference genome with both minimap2 and NGMLR and subsampled to obtain coverages ranging from 5x to 30x. We applied GASOLINE somatic mode on these synthetic sample pairs and compared its performance with Sniffles2, cuteSV and SVIM, (we applied these tools to each sample separately and only kept SVs with no supporting reads in the control sample). For every sequencing coverage, p-value threshold, aligner and variant size, GASOLINE performed better than the other methods which obtained higher amounts of false positive calls and therefore lower precision (Figure 3.6, panels a-d).

Taken together, these results demonstrate that GASOLINE enables accurate detection of somatic SVs from long-read sequencing data, achieving reliable performance across different sequencing technologies, coverage levels, and variant classes. By integrating read-level clustering, statistical testing, and matched-sample comparison, the method provides a robust and reproducible framework for the identification of tumor-specific structural rearrangements, even in heterogeneous or subclonal cancer genomes.

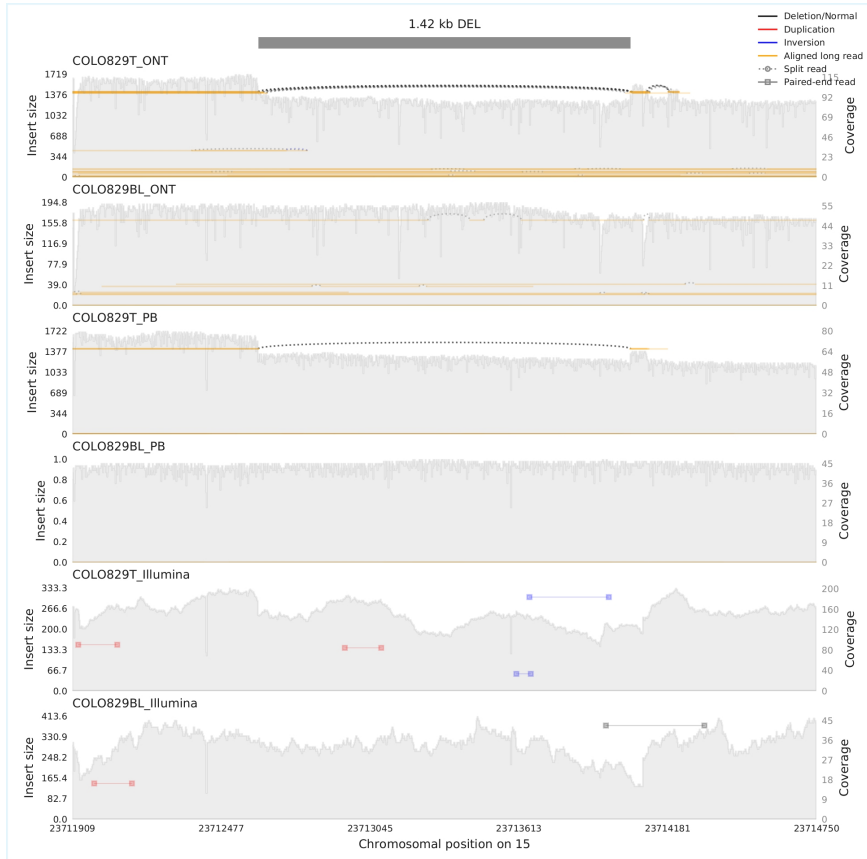


Figure 3.7: The figure shows alignment tracks overlapping one somatic deletion in the COLO829 dataset identified by GASOLINE and uncalled by other tools. The samplots show sequencing reads from ONT, PacBio and Illumina for COLO829 and COLO829BL, which support a somatic, subclonal loss of coverage missed by other methods.

3.5 Materials and Methods

3.5.1 Simulation of synthetic datasets with PBSIM2

We generated synthetic long-read sequencing datasets using PBSIM2, a simulator that generates reads by randomly sampling from a reference sequence and introducing substitutions, insertions, and deletions with a user-defined frequency to mimic long-read data error rates, while also allowing control over the read length distribution and total sequencing coverage.

We simulated datasets with an average length of 10 kb and error rate of 10% to replicate the characteristics of ONT and PacBio reads, with total coverages ranging from 5x to 30x (specifically 5x, 10x, 15x, 20x, 25x, and 30x). All major SV subtypes (deletions, insertions, duplications, inversions, and translocations) were simulated in both homozygous and heterozygous states. To this end, we modified a 5 Mb region of chromosome 1 of the hg19 reference genome (coordinates 1:5,000,001–10,000,000) by introducing variants deletions, insertions, duplications and inversions of increasing size, ranging from 50 b to 5 kb (50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, and 5000 b).

Translocations were simulated by generating two distinct reference sequences from chromosome 1 (regions 1:5,000,001–10,000,000 and 1:10,000,001–15,000,000) and applying PBSIM2 to both. All simulated reads were aligned to the corresponding 5 Mb reference using two long-read mappers, minimap2 and NGMLR.

To extend the benchmarking to more complex rearrangements, we additionally simulated 3000 rearrangements (inversions, duplications, and translocations) in the 500 b - 30 kb range with SURVIVOR [70] by creating a modified reference genome. Synthetic ONT and PacBio reads were then produced from both the modified and unmodified reference genomes at sequencing depths of 2.5x, 5x, 7.5x, 10x, 12.5x, and 15x using PBSIM2. To simulate heterozygosity, reads from the modified and unmodified genomes were mixed prior to alignment with minimap2 and NGMLR.

3.5.2 Comparison with existing SV detection tools

For germline variant analyses, Sniffles2 and SVIM were used with their default settings, while cuteSV was run with the parameters recommended by the developers for ONT data :


```
(--max_cluster_bias_INS=100, --diff_ratio_merging_INS=0.3,
--max_cluster_bias_DEL=100, --diff_ratio_merging_DEL=0.3)
and for PacBio data
```

```
(--max_cluster_bias_INS=100, --diff_ratio_merging_INS=0.3,
--max_cluster_bias_DEL=200, --diff_ratio_merging_DEL=0.5).
```

Somatic variant calling was performed on paired test-control datasets by running each tool independently on both samples and removing all SVs that showed support in the control. The presence of supporting signatures was evaluated using a reciprocal overlap greater than 0.5.

Precision (or Positive Predictive Value), recall (or True Positive Rate), and F1-score were computed as follows:

- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

where TP, FP, and FN indicate true positives, false positives, and false negatives, respectively. Results were further stratified by variant size (< 500 b and > 500 b) and/or genotype (heterozygous vs. homozygous). An SV was considered a true positive when its reciprocal overlap with a validated event was $\geq 50\%$. In the somatic context, statistical significance was evaluated using Fisher's exact test.

3.5.3 NA24385 dataset (GIAB)

Long-read sequencing data for the Ashkenazim individual NA24385 were retrieved from the Genome in a Bottle (GIAB) repository (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/). ONT reads (FASTQ format) and PacBio reads (FASTA format) were aligned to the hg19 human reference genome with minimap2 and NGMLR and downsampled with seqtk (<https://github.com/lh3/seqtk>) to generate 5x, 10x, 15x, 20x, 25x, and 30x subsets.

The GIAB consortium provided a high-confidence benchmark of germline SVs for NA24385, integrating results from short-, linked-, and long-read sequencing together with optical mapping. This reference callset comprises 12,745 variants, including 7,281 insertions and 5,464 deletions. The validated dataset was downloaded from https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/.

gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/.

3.5.4 COLO829 dataset

The COLO829 melanoma cell line and its matched normal counterpart COLO829BL were used to assess GASOLINE’s capability to detect somatic SVs. Both lines were previously characterized by Valle-Inclan et al. using a combination of Illumina HiSeq XTen, ONT, PacBio, 10x Genomics, and Bionano optical mapping. From these multi-technology data, a validated somatic SV truth set was compiled, including 68 high-confidence events (38 deletions, 3 insertions, 7 duplications, 7 inversions, and 13 translocations), available at <https://doi.org/10.5281/zenodo.3988185>.

Whole-genome sequencing data for COLO829 and COLO829BL were downloaded from EGA project PRJEB27698 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB27698>). ONT and PacBio reads (FASTQ format) were aligned to the hg19 reference genome using minimap2, obtaining mean coverages of approximately 60x and 50x, respectively.

3.6 Conclusions

The work presented in this chapter describes the development and validation of GASOLINE, a novel computational tool for the identification of SVs from long-read sequencing data. The motivation behind GASOLINE arises from the limitations that still affect gold-standard approaches for this task: existing algorithms often struggle to achieve high accuracy and breakpoint precision across different SV classes, particularly in the presence of sequencing noise, alignment uncertainty, or low coverage conditions. Our approach introduces a refined clustering framework based on the concept of NRO, designed to improve the grouping of SV signatures and reduce both fragmentation and redundancy in variant detection.

A first extensive validation was carried out using synthetic datasets for all major SV classes over a wide range of sizes (50 bp - 5 kb) and sequencing depths (5x-30x), under both homozygous and heterozygous conditions. The results demonstrated that GASOLINE maintains stable performance across coverage levels and sequencing technologies, showing particularly high precision in clustering true variant signatures, even in noisy or low-depth

datasets. When benchmarked against state-of-the-art SV callers such as Sniffles2, cuteSV, and SVIM, our tool achieved comparable or superior performance, especially for inversions, duplications, and translocations, variants that are typically difficult to resolve. These findings support the robustness of the NRO-based clustering strategy as a general approach for long-read SV discovery.

The analysis of real data further confirmed the potential of GASOLINE. Using the well-characterized GIAB dataset for the NA24385 sample, we demonstrated that our method accurately identifies germline insertions and deletions across a broad spectrum of variant sizes, achieving the highest F1-scores for smaller events (< 500 bp), where alignment ambiguity and read errors typically have a stronger impact. Importantly, these results were consistent across both ONT and PacBio data and independent of the aligner used, highlighting the general applicability of the algorithm.

Beyond germline analysis, GASOLINE was designed to handle the detection of somatic variants from matched sample pairs a feature rarely available in existing long-read SV callers. In cancer genomics and other mosaic contexts, somatic SV detection requires distinguishing true tumor-specific rearrangements from germline background events. To this end, GASOLINE first applies the germline detection module to the tumor (or test) sample and then searches for corresponding signatures in the matched normal sample, quantifying their similarity through the NRO metric. SVs lacking support in the control are statistically tested using Fisher’s exact test to assess the significance of the imbalance in supporting reads between the two conditions. This framework provides a principled way to identify somatic rearrangements while accounting for sequencing depth and allelic fraction.

We assessed this functionality using the COLO829 melanoma cell-line dataset and its normal counterpart COLO829BL, a gold-standard resource for somatic SV benchmarking. GASOLINE accurately recovered the majority of validated somatic events, demonstrating high concordance with the established truth set and confirming its suitability for complex genomic samples such as tumors. The method also detected additional rearrangements supported by ONT reads but absent from short-read or optical mapping callsets, reflecting the increased sensitivity of long-read sequencing for low-frequency or repetitive-region variants.

Taken together, these results highlight the versatility and reliability of GASOLINE as a tool for comprehensive SV discovery from long-read se-

quencing data. The introduction of the NRO-based clustering approach substantially improves the consistency and precision of variant detection, particularly for complex and heterogenous genomic rearrangements. Moreover, the integration of a dedicated somatic-calling module extends its applicability beyond germline genomics, enabling the exploration of tumor evolution, mosaicism, and other dynamic genomic processes.

3.6.1 Future developments

After publication, we have been working on additional functionalities and performance improvements for GASOLINE. Specifically, we are evaluating modifications to the clustering algorithm in order to concurrently analyze signatures for different SV types and improve accuracy in highly rearranged or complex regions, which are often uncalled with standard approaches.

Moreover, we integrated an annotation module that enables the direct identification of genes affected by each structural variant, thereby facilitating the evaluation of their potential functional impact. Specifically, for each SV, the annotation reports genes involved in copy number gains, losses, or truncations, and provides a pathogenicity score derived from the ACMG guideline criteria.

Chapter 4

Differential Methylation detection

*DNA methylation is a dynamic regulator of genome function, yet standard analytical frameworks often reduce this complex landscape to simple average methylation levels (β -values), thereby discarding the rich information contained in cell-to-cell variability. Nanopore sequencing, by detecting methylation states along single native DNA molecules, offers the unique opportunity to resolve epialleles. In this chapter we describe PoreMeth2, our novel computational method specifically designed for the detection of Differentially Methylated Regions between evolutionarily related sample pairs, such as longitudinal time-points or matched healthy-cancer tissues. Unlike conventional tools that rely solely on mean methylation differences, PoreMeth2 introduces a bivariate segmentation approach that simultaneously integrates changes in methylation frequency ($\Delta\beta$) with changes in differential epiallelic entropy (ΔS). By coupling these two dimensions, PoreMeth2 can provide a deeper, mechanistic understanding of epigenetic variation, revealing regulatory dynamics that remain invisible to standard univariate analyses.*¹

¹This chapter has been published as "PoreMeth2 for decoding the evolution of methylation alterations with nanopore sequencing" in Genome Research, 2025 [73]

4.1 Background

The study of DNA methylation has been a central focus of epigenetics for several decades, given its pivotal role in regulating gene expression, maintaining genome integrity, and driving both physiological and pathological processes.

4.1.1 Early assays for methylation detection

Historically, the first experimental strategies to investigate DNA methylation patterns were based on the use of methylation-sensitive restriction enzymes (MSREs), which enabled indirect detection of methylated cytosines at specific recognition sites. Enzymes such as HpaII and MspI, which recognize the same palindromic sequence (CCGG) but differ in their sensitivity to methylation, were widely employed to assess the methylation status of discrete genomic regions. In these assays, genomic DNA is digested with one or both enzymes and analyzed via Southern blot or PCR amplification to infer methylation at the enzyme's recognition sites [74], [75]. The differential cleavage patterns provided the first functional readout of methylation variability between samples. While these early approaches laid the foundation for epigenetic analysis, they were inherently limited by their restriction-site dependence, confining methylation profiling to a small subset of CpGs within enzyme recognition motifs and preventing genome-wide characterization.

To expand upon these locus-specific methods, several hybrid strategies were later developed to combine restriction enzyme digestion with higher-throughput detection technologies. Methylation-sensitive amplified polymorphism (MSAP) and methylation-sensitive restriction fingerprinting (MSRF), for instance, enabled the analysis of methylation polymorphisms across hundreds of loci by coupling restriction digestion to PCR and gel electrophoresis. More recently, methylation-sensitive restriction enzyme sequencing (MRE-seq) adapted this principle to next-generation sequencing (NGS) platforms, allowing genome-wide profiling of unmethylated CpG sites. In MRE-seq, DNA is digested with a panel of methylation-sensitive restriction enzymes, and the resulting fragments are sequenced to identify hypomethylated regions with base-pair precision [76]. However, even this advancement remained inherently biased toward enzyme recognition sites and could not detect fully methylated or non-CpG methylated regions, thus providing an incomplete

picture of the methylome.

In parallel, antibody- or affinity-based enrichment methods emerged as alternative strategies for global methylation profiling. Techniques such as methylated DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain-based capture (MBD-seq) exploit the ability of antibodies or methyl-binding proteins to selectively bind methylated DNA fragments, which are subsequently purified and analyzed by microarray or sequencing [77], [78]. These methods offered the first truly genome-wide view of methylation landscapes, but their resolution was constrained by fragment size (typically 100-500 bp) and by the semi-quantitative nature of enrichment signals, which precluded precise estimation of methylation levels at single CpG resolution. Moreover, enrichment-based approaches exhibit preferential capture of CpG-rich regions, thereby underrepresenting sparsely methylated or non-CpG contexts.

Collectively, these early enzymatic and affinity-based techniques represented critical milestones in the evolution of epigenomic research. They transformed DNA methylation from a locus-specific curiosity into a genome-wide phenomenon, setting the stage for the advent of bisulfite conversion-based technologies that would later enable single-base resolution and quantitative methylation profiling across entire genomes.

4.1.2 Bisulfite-based methylation sequencing

The advent of bisulfite sequencing in the 1990s represented a transformative step, allowing single-base resolution analysis of DNA methylation. Sodium bisulfite treatment converts unmethylated cytosines into uracil while leaving methylated cytosines unchanged; after PCR and sequencing these uracils are read as thymine, allowing direct inference of methylation at single cytosines [79]. This technique laid the foundation for gold-standard methods such as whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS), which have been extensively applied to profile DNA methylation across diverse biological systems.

WGBS applies bisulfite conversion to fragmented genomic DNA followed by high-throughput sequencing, yielding genome-wide, base-level methylation maps. The approach provides the most comprehensive and unbiased methylome profiles possible with bisulfite chemistry and was instrumental in revealing previously unappreciated features of mammalian methylomes,

including extensive cell-type specific patterns and non-CpG methylation in embryonic stem cells [80].

RRBS was developed as a cost-effective compromise that enriches for CpG-dense regions while retaining base resolution. RRBS uses methylation-insensitive restriction enzymes (commonly MspI) and size selection to generate a reproducible subset of the genome highly enriched for promoters and CpG islands; these fragments are then treated with bisulfite conversion and sequencing. RRBS interrogates only a fraction of all CpGs (typically focused on promoters, CpG islands and other CpG-rich elements), but at a markedly reduced sequencing cost compared to WGBS, making it an attractive option for studies where promoter-centric methylation is of primary interest [81].

Despite their widespread use, bisulfite-based methods suffer from several drawbacks, including harsh chemical treatment leading to DNA degradation [82], high sequencing costs due to the need for deep coverage [83], and the inability to distinguish between different cytosine modifications, such as 5-methylcytosine and 5-hydroxymethylcytosine [84].

4.1.3 Methylation Arrays

In parallel, array-based platforms such as Illumina Infinium arrays and EPIC arrays offered a cost-effective alternative for large-scale methylation studies, enabling standardized, high-throughput interrogation of hundreds of thousands of CpG sites across the genome, and became widely adopted in population-scale and clinical studies.

The Infinium technology is based on site-specific hybridization of bisulfite-converted DNA to oligonucleotide probes immobilized on microscopic beads. Each probe is designed to target a single CpG site and to discriminate between the methylated and unmethylated forms of the cytosine through single-base extension and fluorescent labeling. In the most widely used designs, two distinct probe architectures are implemented: Infinium I, which uses two separate probes per CpG site (one recognizing the methylated and one the unmethylated sequence), and Infinium II, which relies on a single probe and differentiates methylation states by color-channel detection after a single-base extension step. These two probe architectures allow precise quantification of methylation levels, expressed as a β -value ranging from 0 (fully unmethylated) to 1 (fully methylated) [85].

The first commercial version of the array, the Infinium HumanMethylation27 BeadChip, quantified approximately 27,000 CpG sites mainly located in promoter-associated CpG islands of $\sim 14,000$ genes [86]. Its successor, the HumanMethylation450 BeadChip, released in 2011, expanded coverage to over 485,000 CpG sites, including promoter, gene body, intergenic, and enhancer-associated loci [85]. The platform was later upgraded to the Infinium MethylationEPIC BeadChip (EPIC v1), covering over 850,000 CpGs by incorporating enhancer elements identified in the ENCODE and FANTOM5 projects [87]. Recently, the newest iteration, the EPIC v2 array, has further extended this coverage to over 935,000 CpG sites, improving the representation of distal regulatory regions and addressing technical dropouts present in earlier designs **noguera2023validation**.

The strengths of array-based methylation assays lie in their high reproducibility, low per-sample cost, and compatibility with degraded DNA, such as that derived from formalin-fixed paraffin-embedded (FFPE) clinical samples. However, the probe-based nature of arrays introduces certain limitations. Coverage is inherently restricted to the CpG sites represented on the chip, which, despite being genome-wide in distribution, cover less than 3-5% of all CpGs in the human genome. Moreover, probe cross-reactivity, polymorphisms within probe-binding regions, and sequence context effects can lead to spurious signals or reduced accuracy, particularly in repetitive regions or CpG-poor contexts. Furthermore, because they rely on bisulfite conversion, arrays share the same chemical limitations as sequencing-based bisulfite methods.

4.1.4 Direct long-read sequencing

More recently, the advent of long-read sequencing technologies has fundamentally transformed the field of epigenetic analysis by enabling the direct detection of base modifications from native DNA without chemical conversion or amplification. As described in section 1.2.3, these platforms read individual DNA molecules in real time, capturing not only the sequence information but also modification-dependent signals (ionic current shifts in ONT sequencing or altered interpulse durations in PacBio's SMRT sequencing). As a result, methylation and other epigenetic modifications can be profiled simultaneously with nucleotide sequences on the same molecule [88], [89].

The ability to analyze native, unamplified DNA confers several major advantages over bisulfite-based methods. First, the chemical integrity of DNA is preserved, avoiding the degradation and coverage bias introduced by bisulfite treatment, which is particularly detrimental in GC-rich or repetitive regions. Second, long continuous reads, often exceeding tens or hundreds of kilobases, enable the direct measurement of methylation patterns across large genomic domains, spanning entire promoters, enhancers, or even complete gene loci. This continuity allows to study the relationship between neighboring CpGs along the same molecule, and facilitates the reconstruction of haplotype-resolved methylomes [90].

Furthermore, the single-molecule nature of long-read data provides access to epiallelic heterogeneity, allowing the identification of mixed methylation states that reflect underlying clonal diversity or subpopulation-specific regulation, features that are averaged out in conventional short-read or array-based approaches. This capability is particularly relevant in complex systems such as tumors, where the coexistence of genetically and epigenetically distinct subclones underlies disease progression and treatment resistance.

From a technical standpoint, long-read platforms also expand the range of detectable base modifications. Unlike bisulfite conversion, which confounds 5-methylcytosine (5mC) with 5-hydroxymethylcytosine (5hmC), signal-based detection preserves the native chemical identity of the base and can, in principle, distinguish multiple cytosine and adenine modifications (e.g., 5mC, 5hmC, 6mA) when appropriate machine-learning models are applied to the raw signal [88].

Finally, despite the typically higher error rates with respect to short-read sequencing technologies, continuous advances in sequencing chemistry and basecalling algorithms have progressively enhanced the quantitative accuracy and throughput of methylation detection, making it feasible to perform whole-genome methylation profiling at population or clinical scale. These developments have positioned long-read sequencing as the most comprehensive approach currently available for the integrated characterization of the epigenome.

4.2 State of the art

One of the main goals of DNA methylation analysis is the identification of differentially methylated regions (DMRs), contiguous genomic segments

Table 4.1: Summary of the main currently used methods for DNA methylation profiling

Method	Principle	Strengths	Limitations
Bisulfite sequencing (WGBS, RRBS)	Sodium bisulfite converts unmethylated cytosines to uracil; sequencing reveals methylation status at single-base resolution	Base-pair resolution, genome-wide profiling	DNA degradation, high sequencing cost, inability to distinguish 5mC vs 5hmC
Microarrays (e.g. 450K, EPIC)	Hybridization of bisulfite-treated DNA to arrays with CpG-specific probes	High-throughput, cost-effective, standardized platforms	Limited to predefined CpG sites, incomplete genome coverage
Long-read sequencing (PacBio, ONT)	Direct detection of base modifications in native DNA molecules during sequencing	Single-molecule resolution, allele-specific methylation, long-range phasing, no chemical conversion needed	Higher error rates, specialized computational tools required, still costlier than arrays

that exhibit significant differences in methylation levels between two biological conditions. These regions may serve as molecular signatures of altered epigenetic regulation, often reflecting changes in gene expression, chromatin accessibility, or broader cellular states. DMR analysis thus represents a cornerstone in both fundamental and translational epigenomics, providing insights into mechanisms of development, differentiation, environmental response, and disease pathogenesis.

In nanopore sequencing, methylation is directly inferred from deviations in the ionic current signal produced as DNA translocates through the nanopore. Several tools such as Nanopolish [91], Tombo [92], DeepSignal [93] in addition to ONT proprietary tools such as Guppy and the more recent Dorado implement statistical or machine learning models to convert these current deviations into per-read base modification calls (typically reported in BAM or BEDMethyl format). These programs therefore represent the first analytical step in Nanopore-based methylation workflows, producing base-level estimates of methylation probability or frequency at CpG sites.

In PacBio SMRT sequencing, methylation is detected by monitoring the kinetics of the DNA polymerase during single-molecule real-time synthesis. The presence of epigenetic marks, such as 5-methylcytosine (5mC), induces characteristic pauses or delays in the enzyme’s translocation speed, primarily measurable as alterations in the Interpulse Duration (IPD)—the time interval between consecutive fluorescent pulses. To decode these kinetic signatures, specialized tools such as Primrose [42] and Jasmine, as well

as open-source alternatives like ccsmeth [43], employ deep learning models trained on HiFi reads. These algorithms generate per-read methylation probabilities that are embedded directly within the sequence alignment files, thereby providing the foundational single-molecule epigenetic data required for downstream DMR analysis.

However, few methods for the identification of DMRs from long read data have been developed, and they are mainly adaptations of algorithms designed for short-read sequencing data. For instance, NanoMethPhase [94] includes a function for DMR detection that is a wrapper of the short-read based DSS library [95]. Other methods newly developed for nanopore sequencing data include the Modkit (<https://github.com/nanoporetech/modkit>) DMR function and PycoMeth’s [96] Meth.Comp, which however simply compare methylation frequency levels across samples without taking long range epiallelic information into account. Moreover, PycoMeth reports differential methylation frequency on user-provided query regions, rather than at whole genome level, and it does not include segmentation steps for DMR identification, while Modkit provides Hidden Markov Model (HMM) based segmentation.

Similarly, the library PoreMeth released by our laboratory in 2023 [97] performs differential methylation frequency segmentation by means of a Shifting Level Model (SLM) [98]. One of the main motivations behind this development was the recognition that long-read methylation data exhibit a markedly different CpG organization compared to short-read bisulfite sequencing datasets. In conventional WGBS, CpG sites are densely sampled, and DMRs are typically detected as contiguous stretches of CpGs displaying consistent methylation shifts, obtained through segmentation of $\Delta\beta$ values along spatially ordered cytosines.

However, this assumption does not hold for long-read data, which naturally encompass extended genomic intervals with very few CpG sites, interspersed with short CpG-rich regions such as CGIs. Methods developed for WGBS thus tend to over-segment sparse regions or miss extended low-density DMRs, leading to an inaccurate representation of methylation changes in long-read datasets.

The SLM implemented in PoreMeth explicitly incorporates inter-CpG genomic distances into its likelihood formulation. By doing so, the model dynamically adapts to local CpG density, merging sparse CpG signals into coherent regions when appropriate and maintaining fine-scale resolution in

CpG-rich contexts. This distance-aware formulation provides a more faithful reconstruction of the methylation landscape captured by long reads, effectively extending DMR detection to genomic territories that were previously inaccessible to short-read methods.

Despite these improvements, PoreMeth (as most existing tools) was primarily focused on differences in methylation frequency ($\Delta\beta$) between experimental conditions, without explicitly accounting for other sources of epigenetic variability that can be captured through long-read sequencing. In particular, it did not model within-sample heterogeneity or epiallelic diversity, which may provide complementary insights into the regulatory and clonal dynamics underlying methylation changes.

To address these aspects, we developed PoreMeth2 [73], an extension of the original framework that incorporates methylation entropy (S) as an additional parameter alongside methylation frequency. This quantity reflects the degree of variability in methylation states among individual reads within a region and can therefore capture alterations in the distribution and organization of epialleles, rather than only in their mean methylation levels. In PoreMeth2, these two complementary dimensions are integrated within a Bivariate Shifting Level Model (BiSLM), which allows the joint segmentation of $\Delta\beta$ and ΔS signals across the genome.

4.3 Results

4.3.1 Methylation Entropy

Classical frequency-based approaches reduce methylation to a proportion (the fraction of methylated CpGs within a genomic region), and they implicitly assume that all reads contribute equally to this average and that the underlying population is uniform. For instance, for a region with $\beta = 0.8$, such a value may arise from two entirely different scenarios: one in which each molecule is approximately 80% methylated, and another in which half the molecules are completely methylated and the other half completely unmethylated. Both yield the same mean frequency but correspond to profoundly different biological states (one homogeneous, the other heterogeneous).

The central innovation introduced by PoreMeth2 is the explicit integration of methylation entropy as a quantitative descriptor of epigenetic heterogeneity. The concept of entropy originates from information theory, where it

represents a measure of uncertainty or disorder within a system. Applied to DNA methylation, it quantifies the diversity of methylation states observed across single molecules (and therefore different cells in the sequenced sample) in a given genomic region. Intuitively, entropy captures how variable or consistent the methylation landscape is across the population of DNA molecules sequenced.

Formally, methylation entropy (S) for a region can be computed using Shannon's definition [99]:

$$S = \frac{e}{b} \sum \left(-\frac{n_i}{N} \text{Log} \frac{n_i}{N} \right) \quad (4.1)$$

where:

- N is the number of reads covering the region;
- n_i represents the probability of observing a specific epiallelic pattern (i.e., a unique binary configuration of methylated and unmethylated sites along a read) among all reads covering that region;
- e is entropy for code bit;
- b is the number of CpGs in the region.

For the computation of methylation entropy, a sliding window approach is applied across the ordered CpG positions. Starting from the first group of b consecutive CpGs, S is calculated and the window is then shifted by one CpG at a time to evaluate the next interval. For each CpG position j , S is thus computed on the set of epialleles spanning from CpG_j to CpG_{j+b-1} . Reads containing uncalled methylation states within the considered b -CpG window are excluded from the entropy estimation for that specific interval.

S can take value between 0 and 1: specifically, when the probability distribution is dominated by a single configuration (e.g., all reads share the same pattern) S is close to 0; when multiple configurations coexist with comparable frequencies, S increases (Figure 4.1A). This property makes entropy a powerful measure to assess the internal organization of the epigenome beyond simple averages.

This metric effectively translates the combinatorial complexity of methylation states into a single numerical value, enabling statistical comparison between samples. When compared between two conditions, changes in entropy

(ΔS) provide insight into the underlying dynamics of epigenetic regulation beyond mean methylation shifts ($\Delta\beta$).

A decrease in entropy ($\Delta S < 0$) reflects a reduction in epiallelic diversity, indicating that one specific methylation pattern has become predominant within the population of molecules. This condition can be interpreted as a form of epiallelic selection, where a particular methylation configuration (either hypermethylated or hypomethylated) confers a selective advantage or becomes fixed as a consequence of clonal expansion. Such patterns are often observed in stable or differentiated cellular states, or in tumors where selective pressure favors the emergence of dominant epigenetic clones.

Conversely, an increase in entropy ($\Delta S > 0$) denotes a gain in epiallelic diversity, suggesting the appearance of new or heterogeneous methylation configurations within the region. This condition may reflect stochastic or environmentally induced changes, where methylation patterns become more variable due to partial remodeling, transient deregulation, or stress-related epigenetic instability. By jointly considering both $\Delta\beta$ and ΔS , it becomes possible to discriminate between directed and selective methylation changes (low-entropy, high-consistency shifts) and random or unstable alterations (high-entropy, heterogeneous shifts). This dual perspective provides a more complete and biologically meaningful view of differential methylation, capturing not only the direction of change but also the degree of order or disorder underlying it.

Specifically, with our method DMRs are categorized as one of six possible changes based on the combination of $\Delta\beta$ and ΔS (Figure 4.1B):

- $\Delta\beta > 0$ and $\Delta S > 0$: random hypermethylation (Figure 4.1A1)
- $\Delta\beta < 0$ and $\Delta S < 0$: selective hypomethylation (Figure 4.1A2)
- $\Delta\beta > 0$ and $\Delta S \sim 0$: hypermethylation with no entropy change (Figure 4.1A3)
- $\Delta\beta < 0$ and $\Delta S \sim 0$: hypomethylation with no entropy change (Figure 4.1A4)
- $\Delta\beta > 0$ and $\Delta S < 0$: selective hypermethylation (Figure 4.1A5)
- $\Delta\beta < 0$ and $\Delta S > 0$: random hypomethylation (Figure 4.1A6)

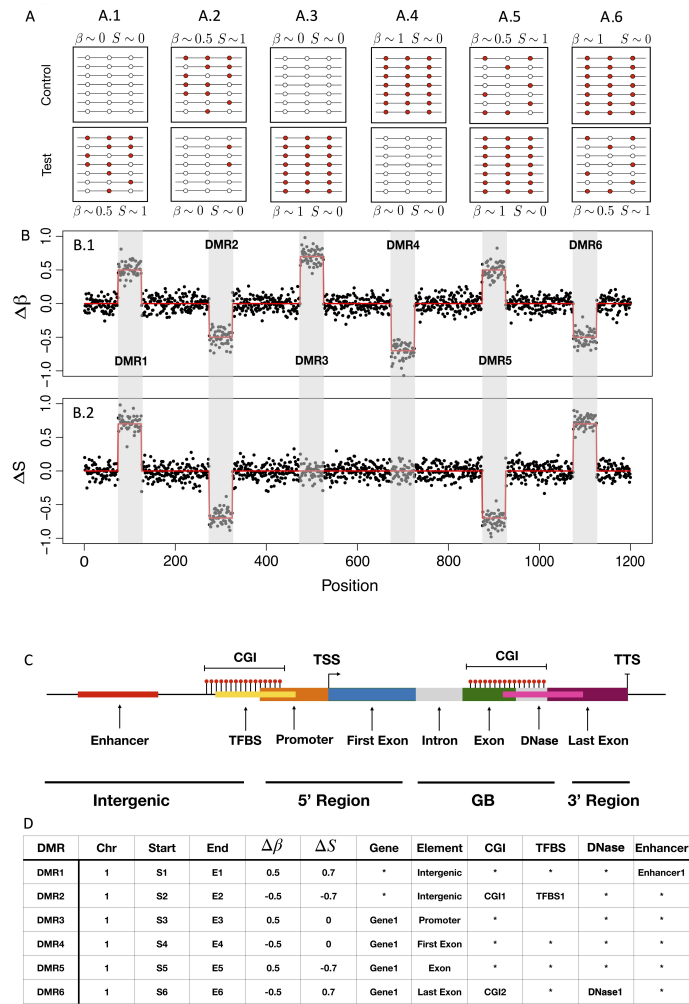


Figure 4.1: Panels A and B display the possible combinations of $\Delta\beta$ and ΔS values, defining the six DMR categories described in section 4.3.1. Panel C shows the scheme behind our annotation and prioritization framework, underlining the complexity of possible overlaps among genic and regulatory elements. Panel D exemplifies the tabular output structure of the annotation scheme, with one or multiple rows for each DMR (one for every overlapping genic element).

4.3.2 BiSLM

In order to perform the co-segmentation of $\Delta\beta$ and ΔS , we introduced in PoreMeth2 the Bivariate Shifting Level Model (BiSLM). BiSLM builds on the mathematical foundation of Hidden Markov Models (HMMs), which are widely used in computational biology for problems where observed data are generated by underlying hidden states with probabilistic transitions. The key intuition behind BiSLM is to model methylation as a bivariate signal, where one dimension corresponds to $\Delta\beta$ and the second dimension represents ΔS . The "shifting level" component reflects the fact that the underlying methylation distribution can change across genomic coordinates, and the algorithm is structured to detect such changes by transitioning between latent states representing different methylation regimes.

From a technical perspective, BiSLM operates by analyzing the methylation calls of individual reads aligned to a reference genome and segmenting the genome into candidate regions. For each window, two complementary statistics are computed:

- Methylation frequency difference ($\Delta\beta$) between two samples, representing the classical measure used by most DMR callers. This captures whether the proportion of methylated CpGs significantly differs across conditions.
- Methylation entropy difference (ΔS), which quantifies the diversity of methylation patterns observed across molecules within each group. Entropy is calculated according to Shannon's information theory, where higher entropy indicates greater variability in epiallelic states. By comparing entropy across groups, BiSLM detects not only shifts in average methylation but also reorganizations in epigenetic heterogeneity.

The HMM structure of BiSLM enables the algorithm to model dependencies between adjacent CpG sites and to account for the fact that methylation states along a DNA molecule are not independent but exhibit local correlation. This makes the model sensitive to the detection of contiguous regions rather than isolated sites, improving robustness to noise and sequencing errors. Moreover, the probabilistic framework allows BiSLM to explicitly estimate the likelihood that a given region belongs to a differentially methylated state versus a background state, which provides a natural mechanism for statistical testing.

From a biological perspective, the introduction of entropy into DMR analysis provides a formal way to capture epigenetic heterogeneity. In development, changes in entropy can reflect processes of lineage commitment and the homogenization of epigenetic states within differentiating cell populations. In disease, particularly in cancer, entropy often increases as a consequence of genomic instability and the coexistence of multiple tumor subclones. Thus, methylation entropy represents a biologically meaningful dimension that complements traditional analysis.

Finally, by integrating entropy into a statistically rigorous HMM framework, BiSLM provides a powerful approach that is both sensitive and biologically interpretable. The method bridges the gap between single-molecule information provided by nanopore sequencing and population-level comparisons required for biological interpretation. It thereby positions PoreMeth2 as an advance over existing tools, enabling researchers to exploit the full informational content of long-read epigenomic data.

With several tests on simulated methylation data we showed that BiSLM is able to accurately detect DMRs with size spanning from 5 to 50 CpGs with high boundary precision with a coverage of 20x or higher. The same simulated data was analyzed with two other methods for DMR detection, Modkit (natively developed for nanopore data) and BSsmooth (adapted from a bisulfite-sequencing based pipeline), and BiSLM performed better than both algorithms in the identification of DMR boundaries (Figure 4.2), while also integrating ΔS information [73].

Model Description

The algorithm models the bivariate signals $x = (x_1, \dots, x_i, \dots, x_N)$, where $x_i = (\Delta\beta_i, \Delta S_i)$ for N CpG sites, as the sum of an underlying state process and a site-independent error term:

$$x_i = m_i + \epsilon_i \tag{4.2}$$

where:

- x_i is the bivariate observation vector (the measured $\Delta\beta$ and ΔS data) at site i .
- $m_i = (m_{i1}, m_{i2})$ is the unobserved latent vector representing the true mean level of the signal (the hidden state) at site i .

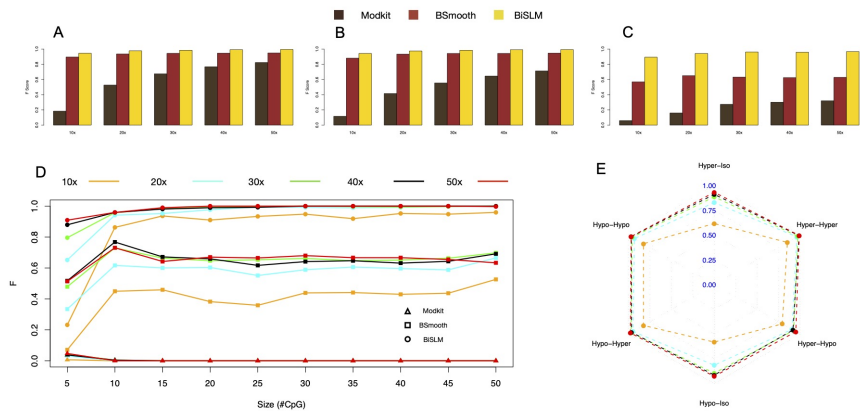


Figure 4.2: Comparative assessment of BiSLM, Modkit, and BSmooth performance on synthetic methylation datasets. Panels A-C show F-measure scores (harmonic mean of precision and recall) for synthetic DMR detection evaluated at reciprocal overlap thresholds of 0.1 (A), 0.5 (B), and 0.9 (C). Panel D displays F-measures obtained by the three methods, plotted as a function of sequencing coverage for varying DMR sizes. The radar plot in E illustrates the Correct Classification Rate (CCR) achieved by the bivariate SLM model in categorizing epiallelic shifts across different coverage depths.

- ϵ_i is the error vector, capturing measurement noise and unmodeled biological variance. This error term is independent across CpG site and it is assumed to follow a bivariate Normal distribution with a zero mean and covariance matrix Σ_ϵ (i.e., $\epsilon_i \sim N(0, \Sigma_\epsilon)$).

The variation of the true mean level m_i from one CpG site to the next is governed by a stochastic process that defines the "shifts" in level:

$$m_i = (1 - z_{i-1}) \cdot m_{i-1} + z_{i-1} \cdot (\mu + \delta_i) \quad (4.3)$$

where:

- z_i is a Bernoulli random variable ($z_i \in \{0, 1\}$) acting as a switch. The probability of a shift is defined as $\eta = \Pr(z_i = 1)$, and the probability of "no-shift" is $(1 - \eta) = \Pr(z_i = 0)$.
- $\mu = (\mu_{\Delta\beta}, \mu_{\Delta S})$ is the global mean vector of the levels.
- δ_i is a random vector drawn from a bivariate Normal distribution $N(0, \Sigma_\mu)$, which introduces variability between different segments.

Conceptually, if $z_{i-1} = 0$ (no shift), the true state remains identical to the previous one ($m_i = m_{i-1}$), indicating the continuation of a segment. If $z_{i-1} = 1$ (a shift occurs), the state m_i is re-sampled from a $N(\mu, \Sigma_\mu)$ distribution, independently of the previous state, thus marking the beginning of a new segment.

The total observed covariance in the data, $Cov(x_i) = \Sigma$, is the sum of the covariances from the two processes:

$$Cov(x_i) = \Sigma = \Sigma_\mu + \Sigma_\epsilon \quad (4.4)$$

The algorithm introduces a parametrization via the parameter $\omega \in [0, 1]$, which balances the impact of these two variance sources:

- $\Sigma_\mu = \omega \cdot \Sigma$ (the between-segment variance)
- $\Sigma_\epsilon = (1-\omega) \cdot \Sigma$ (the within-segment variance, or noise)

The ω parameter acts as a sensitivity control: high ω values imply low-noise data and favor the detection of more segments, whereas low values are suited for noisier data and result in a more conservative segmentation.

A fundamental assumption of the BiSLM is that while the $\Delta\beta$ and ΔS signals are considered independent at both the noise level (Σ_ϵ) and the mean level (Σ_μ), they are correlated by the synchronous switching process z_i . This construct allows the model to identify common shifts, genomic positions where both signals simultaneously change their mean state.

This model, defined with a continuous latent state m_i , poses computational challenges. To overcome this, the BiSLM is formulated as a HMM by discretizing the state space. A Markovian process s_i is introduced, taking values from a finite set of K states, $S = \{1, 2, \dots, K\}$. Each discrete state k is associated with a bivariate mean vector $\mu_k = (\mu_{k1}, \mu_{k2})$, which approximates the continuous variable m_i .

The HMM is then defined by its two main components:

1. Emission Probability ($f_k(x)$): The probability of observing the data x_i , given the hidden state is k . As $\Delta\beta$ and ΔS values are bounded in the range $[-1, 1]$, a truncated bivariate Normal distribution is used:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left[-\frac{1}{2} \left(\frac{x - \mu_k}{\sigma_\epsilon} \right)^2 \right] \cdot \frac{1}{\Phi(\frac{1-\mu_k}{\sigma_\epsilon}) - \Phi(\frac{-1-\mu_k}{\sigma_\epsilon})} \quad (4.5)$$

where Φ is the standard Normal cumulative distribution function.

2. Transition Matrix (P_{jk}): The probability of transitioning from state j to state k . This matrix encodes the logic of the z_i switching process (with shift probability η):

$$P_{jk} = \begin{cases} (1 - \eta) + \eta \cdot g_{jk} & \text{if } j = k \\ \eta \cdot g_{jk} & \text{if } j \neq k \end{cases} \quad (4.6)$$

where g_{jk} represents the probability of reaching the state k following a shift.

Finally, to identify the most likely sequence of hidden states ($s_1, \dots, s_N$) that generated the observed data ($x_1, \dots, x_N$), the algorithm employs the Viterbi algorithm.

The segment boundaries are subsequently identified at the positions i where the optimal state sequence changes (i.e., $s_i \neq s_{i-1}$). The algorithm's primary tuning parameters are ω (balancing noise vs. signal) and η (which regulates the shift probability).

4.3.3 Annotation

Another key innovation of PoreMeth2 with respect to PoreMeth and other available methods for DMR detection, is the integration of an annotation module designed to provide biological and functional interpretation of the detected regions. CpG methylation exerts its regulatory influence in a highly context-dependent manner. Methylation changes occurring in intergenic or repetitive sequences may reflect large-scale chromatin remodeling or transposon silencing, while those overlapping gene promoters, enhancers, or CpG islands can directly affect transcriptional activity and cellular phenotype.

Numerous studies have demonstrated that promoter-associated CpG sites typically display an inverse relationship between methylation and transcription, where hypermethylation represses gene expression by inhibiting transcription factor binding and promoting chromatin compaction. Conversely, gene body methylation is often positively correlated with transcriptional activity, possibly contributing to alternative splicing regulation or suppression of cryptic promoters [12].

In summary, the functional outcome of a methylation event depends not only on its absolute level, but also on its genomic location, the local CpG density, and the regulatory role of the affected element. For this reason, the annotation module of PoreMeth2 was conceived not as a simple post-processing step, but as an essential component of the analytical workflow, aimed at providing a mean for both statistical detection and functional interpretation.

Importantly, genomic and regulatory annotations do not represent discrete, non-overlapping segments along a linear genome, but rather a complex and interlaced landscape of elements that can coexist or partially overlap (Figure 4.1C). A single DMR can intersect one or multiple genic features (such as exons, introns, or promoter regions) belonging to one or even several adjacent genes, while at the same time overlapping regulatory motifs including CpG islands (CGIs), enhancers, transcription factor binding sites (TFBS), or DNase I hypersensitive sites (DHS).

This multilayered organization highlights that functional interpretation of methylation changes cannot rely solely on positional information with respect to gene boundaries, but must instead integrate both structural and regulatory genomic contexts. While a few methods for the annotation of genomic regions to genic and regulatory elements have been developed (such as GoldMine [100], HOMER [101], GenomicDistributios [102], and annotatr

[103]), they do not capture this complex organization. For this reason, the annotation module of PoreMeth2 was designed to perform comprehensive multi-level intersection analyses, providing a unified framework that simultaneously considers genic architecture and regulatory elements. Specifically, for each DMR identified by the BiSLM algorithm, the overlap with individual genic features (such as introns, exons, and promoters) is first computed; for each of these genic elements, the corresponding overlap with different classes of regulatory elements (CGIs, enhancers, TFBSs and DHSs) is then assessed and reported (Figure 4.1D).

In addition to the multilayered genomic and regulatory annotation, PoreMeth2 implements a gene-centered classification strategy based on a feature priority scheme. In this framework, each gene affected by at least one DMR is categorized according to the most functionally relevant genomic feature overlapped by the region. Specifically, genes are assigned to one of three categories: 5 regulatory regions (5'Reg), when the DMR overlaps promoters, 5'UTRs, or first exons; gene bodies (GB), when the overlap occurs within internal introns or exons but excludes 5'Reg and 3'UTR elements; and 3' untranslated regions (3'UTR), when the DMR overlaps the last exon without intersecting upstream regulatory regions.

This hierarchical prioritization reflects the widely accepted central regulatory importance of promoter methylation. According to classical epigenetic models, methylation within promoter-associated CpG islands inhibits transcription factor binding and reduces chromatin accessibility, thus repressing transcriptional initiation. In contrast, unmethylated promoters facilitate the recruitment of transcription factors and RNA polymerase II, enabling active transcription. By assigning precedence to 5Reg regions when a DMR spans multiple genic elements, the annotation procedure ensures that the most likely regulatory consequence of methylation is captured, emphasizing functionally relevant events over structurally neutral ones. Importantly, this feature-priority scheme retains information on the co-occurring overlap with regulatory motifs.

4.3.4 Acute Myeloid Leukemia

To assess the performance and biological relevance of PoreMeth2, we applied it to a set of three Acute Myeloid Leukemia (AML) patients (UD5, UD10 and AML2), sequenced at diagnosis and at relapse after chemotherapy. These

datasets were previously investigated in [97], to test the previous version of PoreMeth2 (PoreMeth), which provided a valuable framework to capture known methylation dynamics associated with chemoresistance. In that earlier study, we showed that relapsed AML samples display widespread gene body hypermethylation at sparse CpG sites, with preferential enrichment in oncogenes, tumor suppressors, and cancer-related signaling pathways. In this work, our aim was instead to establish whether the methylation entropy measure introduced in PoreMeth2 could actually provide insights on the selective nature, and therefore on the functional impact, of the detected DMRs.

For each sample, we calculated β and methylation entropy S (with $b = 3$) on all CpGs throughout the genome and subsequently applied the PoreMeth2 segmentation algorithm to each pair to identify DMRs. The resulting regions were classified according to their methylation frequency change ($\Delta\beta$) and entropy variation (ΔS) in the six categories described in section 4.3.1. Specifically, DMRs were categorized as hypermethylated ($\Delta\beta > 0.2$) or hypomethylated ($\Delta\beta < -0.2$), and further subdivided into hyperentropic ($\Delta S > 0.1$), isoentropic ($-0.1 < \Delta S < 0.1$), or hypoentropic ($\Delta S < -0.1$).

Our analysis identified 3102, 2825, and 2874 DMRs in UD5, UD10, and AML2, respectively, corresponding to approximately 3 Mb of the genome per sample pair (3.18, 3.07 and 3.18). Importantly, the ability of PoreMeth2 to capture subtle methylation and entropy shifts allowed us to extend differential methylation analysis beyond dense CpG regions, such as CpG islands, to include sparsely methylated genomic segments (Figure 4.3B). As also observed in [97], the number of hypermethylated DMRs markedly the hypomethylated in all the samples (1999 vs 1103 for UD5; 2292 vs 533 for UD10; 2575 vs 299 for AML2), indicating a global shift toward hypermethylation at relapse. Both hyper- and hypomethylated DMRs were evenly distributed among the three ΔS classes, suggesting that epiallelic diversity changes occur across the full spectrum of methylation dynamics rather than being confined to specific regions.

To investigate the potential functional implications of the identified regions, DMRs were annotated using the dedicated module of PoreMeth2, integrating both genic and regulatory annotations. Across the three patients, the majority of DMRs overlapped annotated protein-coding genes (62-67%), while approximately 10% mapped to noncoding genes and 30% were located in intergenic regions (Figure 4.3A). Concerning regulatory elements, 90-93%

of DMRs intersected DNase hypersensitive sites (DHSs) and transcription factor binding sites (TFBSs), while a smaller fraction (12-13%) overlapped enhancers. Lastly, as expected based on CpG density distribution, between 18% and 44% of DMRs overlapped CpG islands (CGIs), whereas the remainder were located in low-CpG-density regions, demonstrating that PoreMeth2 can effectively detect methylation differences even in sparsely methylated genomic contexts (Figure 4.3A).

Using the feature-priority gene annotation scheme implemented in PoreMeth2, we further classified protein-coding genes harboring at least one DMR (differentially methylated genes, DMGs) into three major functional categories 5Reg, GB, and 3UTR (based on the criteria explained in the 'Annotation' section). The DMGs identified for each patient were 1779, 1655 and 1873 respectively, most of which were hypermethylated (70-93%). Among these, 25-40% were classified as 5Reg, 60-80% as GB, and approximately 5% as 3UTR in all the sample pairs. Hypermethylated DMGs in 5Reg predominantly overlapped CpG islands (71-86%), while those within gene bodies mainly involved sparse CpG regions (about 55%). In contrast, hypomethylated genes were almost entirely associated with DMRs in low CpG density regions ($\sim 90\%$), independent of their genic position (Figure 4.3C).

When considering epiallelic diversity, most DMGs exhibited hypoeutropic ($\Delta S < -0.1$) or isoentropic ($-0.1 < \Delta S < 0.1$) changes, which are indicative of selective processes leading to the dominance of specific epialleles, rather than random methylation fluctuations. The only notable exception was represented by genes with hypermethylated DMRs in promoter-associated CpG islands, where a predominance of hyperentropic ($\Delta S > 0.1$) changes was observed, suggesting increased heterogeneity at regulatory loci undergoing transcriptional repression (Figure 4.3D).

To further investigate the functional impact of methylation changes, we integrated the DMRs identified by PoreMeth2 with transcriptomic data obtained from RNA-seq experiments performed in triplicate for all six AML samples. We performed differential gene expression analysis between diagnosis and relapse, resulting in the identification of 3997, 4677, and 1759 differentially expressed genes (DEGs) for patients UD10, UD5, and AML2, respectively. These gene sets revealed heterogeneous expression patterns, with variable proportions of up- and downregulated genes across patients, consistent with the strong interindividual variability typically observed in relapsed AML.

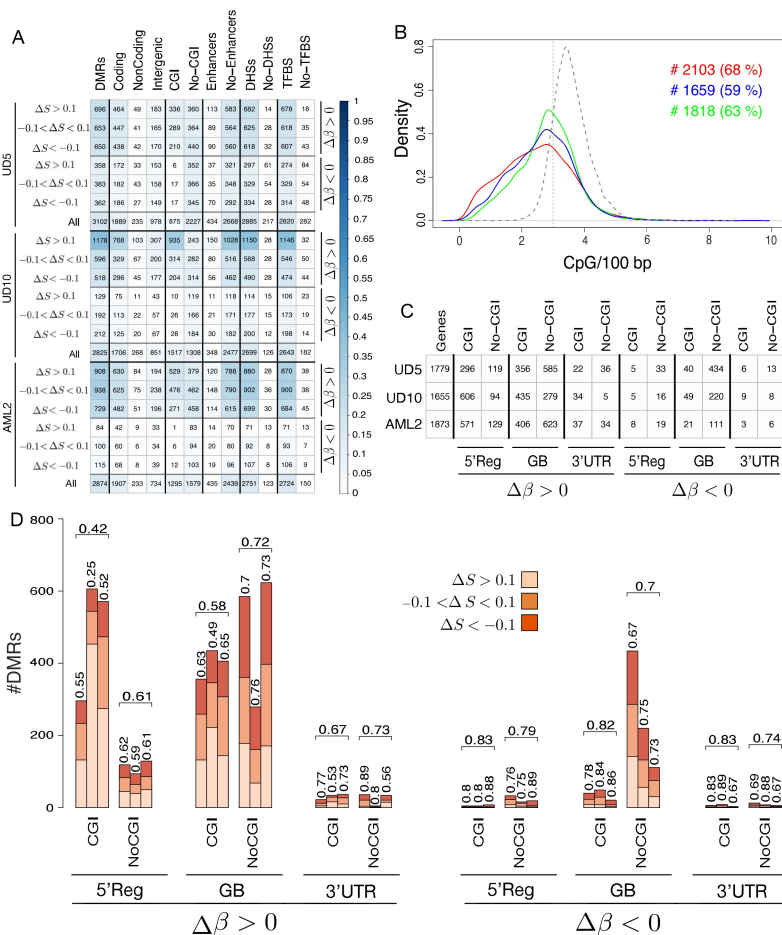


Figure 4.3: Characterization of DMRs in three AML sample pairs. The heatmap in panel A displays the count of DMRs detected by BiSLM, categorized by methylation change ($\Delta\beta$) and entropy variation (ΔS). Rows correspond to DMRs overlapping specific genomic features (coding/non-coding genes, intergenic regions, CGIs, Enhancers, DHSs, and TFBS) or lacking such overlaps (prefix 'No-'). Color intensity indicates the proportional distribution of DMR categories within each sample. Panel B shows CpG density distribution of detected DMRs. Vertical lines mark the resolution limits for WGBS (solid; ≤ 2 CpG/100 bp) and RRBS (dotted; ≤ 3 CpG/100 bp), with adjacent text reporting the count and percentage of DMRs falling below these thresholds. The table in panel C summarizes the number of DMGs harboring DMRs in 5'Reg, GB or 3'UTR, stratified by CGI overlap. Panel D shows a barplot of epiallelic shifts, classified as hyperentropic, hypoentropic or isoentropic, within hyper- and hypomethylated genes. Numbers above bars indicate the percentage of hypo- and isoentropic DM genes, while horizontal brackets denote the average percentages across the three samples.

We then assessed the relationship between DNA methylation alterations and gene expression changes by identifying DMGs that also displayed significant expression variation (DM-DEGs). Across the three patients, the majority of DM-DEGs were associated with DMRs exhibiting hypoeutropic or isoeutropic profiles ($\Delta S < 0.1$), while only a small fraction corresponded to hyperentropic regions ($\Delta S > 0.1$). This pattern suggests that the most transcriptionally relevant methylation changes occur in regions where epiallelic diversity decreases or remains stable, reflecting the clonal selection of specific methylation states rather than random stochastic variation. In contrast, hyperentropic DMRs tended to show weaker or no correspondence with expression changes, implying that they may represent transient or noise-driven epigenetic fluctuations rather than stable regulatory alterations (Figure 4.4A).

When stratifying DMGs according to their genomic context, we observed that genes harboring hypermethylated DMRs within gene bodies and located in sparse CpG regions were significantly enriched for DEGs across all three AML cases (Figure 4.4B). These DMRs were characterized by tumor β values exceeding 0.2 at diagnosis and approaching full methylation ($\beta_{Rel} = 0.7 - 1.0$) at relapse, consistent with a model in which clonal expansion favors cells carrying a dominant, fully methylated epiallele. Conversely, hyperentropic DMRs generally corresponded to loci transitioning from unmethylated to partially methylated states ($\beta_{Tum} = 0 - 0.2$, $\beta_{Rel} = 0.2 - 0.5$), indicating the coexistence of multiple methylation configurations within the tumor population.

We next examined whether the presence and type of DMRs within specific genic features (5 regulatory regions, gene bodies, or 3 UTRs) influenced transcriptional behavior. Hypermethylated DMRs in 5 regulatory regions (particularly those overlapping CGIs) were predominantly associated with transcriptional repression, whereas hypermethylation within gene bodies correlated with either stable or increased expression, in agreement with the dual role of methylation depending on its genomic location. Notably, genes with DMRs in gene bodies at sparse CpG regions, coupled with hypo- or isoeutropic changes, displayed the strongest enrichment among DEGs, underscoring the regulatory relevance of these less CpG-dense domains in modulating transcriptional output.

To explore the biological significance of these DM-DEGs, we performed over-representation analyses (ORA) against curated gene sets representing

major cancer-related categories, including oncogenes and tumor suppressors (COSMIC [104]), transcription factors (TFs; [105]), chemoresistance-associated genes (GEAR [106]), and canonical signaling pathways (KEGG [107]). The ORA results revealed a clear functional stratification according to the epiallelic change category. DM-DEGs with hyperentropic DMRs or with DMRs overlapping CpG islands showed limited enrichment in any cancer-related category, regardless of methylation direction or expression change. In contrast, hypermethylated DMGs exhibiting low or stable entropy variation ($\Delta S < 0.1$) and located within gene bodies at sparse CpG regions were significantly enriched for transcription factors and cancer-related pathways in all three patients (Figure 4.4C). These included genes involved in cellular proliferation, apoptosis resistance and DNA damage response, hallmarks of therapy-resistant AML phenotypes.

Taken together, these findings support the notion that entropy variation provides a valuable dimension for interpreting the functional consequences of DNA methylation changes. Specifically, DMRs characterized by negative or near-zero ΔS values appear to mark regions where selective clonal expansion has led to the fixation of particular epiallelic configurations, thereby exerting a measurable influence on transcription. Conversely, positive ΔS values may represent epigenetic diversification or plasticity, which, while indicative of heterogeneity, do not necessarily correspond to stable regulatory shifts. In this framework, the combination of $\Delta\beta$ and ΔS implemented in PoreMeth2 enables the discrimination between epiallelic selection and random methylation drift, enhancing our ability to link methylation dynamics with functional outcomes in cancer evolution and relapse.

4.3.5 Human Peripheral Glial Cells

To further assess the versatility of PoreMeth2 across different biological and experimental contexts, we applied the tool to a model system of Human Peripheral Glial Cells (HPGCs) subjected to pharmacological stimulation. This dataset was specifically designed to capture short-term methylation dynamics rather than long-term clonal selection, allowing us to evaluate the capacity of our approach to detect subtle regulatory alterations over limited time scales. HPGC cultures were sequenced at baseline (T0) and after 48 hours (T48) of exposure to a G protein-coupled receptor (GPCR) agonist. Each condition was sequenced with high coverage (20–30x) using ONT’s PromethION2 Solo ensuring sufficient read depth and length for reliable

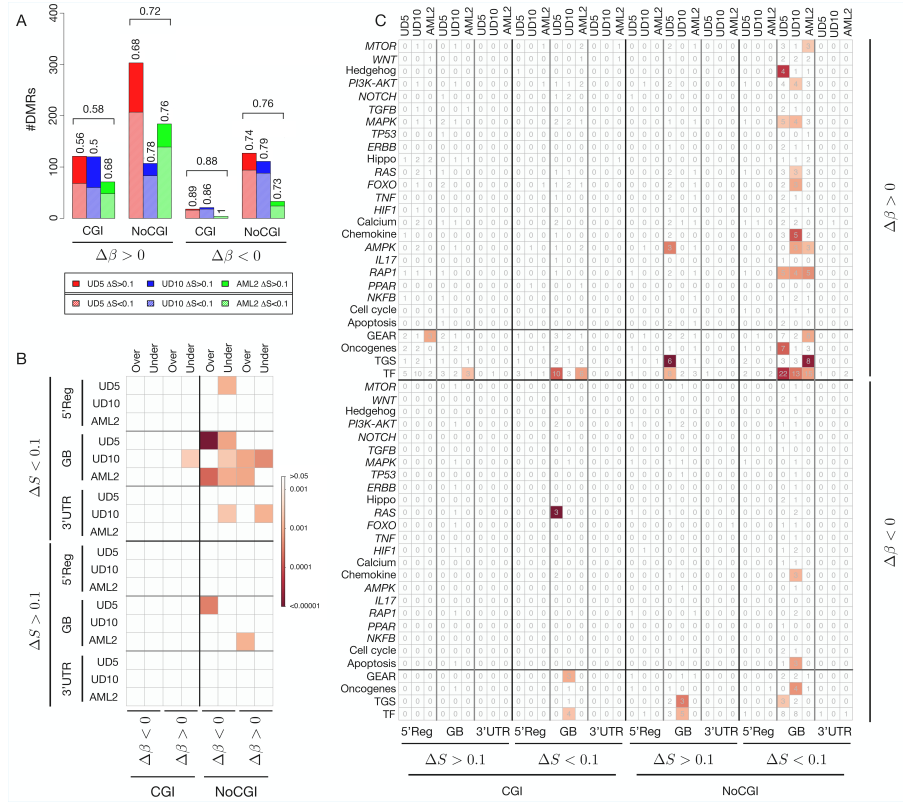


Figure 4.4: Association between DMRs and Differential Expression. Panel A shows the proportion of DM-DEGs characterized by hyperentropic versus iso/hypoentropic DMRs. Data are stratified by methylation status ($\Delta\beta > 0$ or $\Delta\beta < 0$) and CGI overlap (CGI vs. NoCGI). Textured bars specifically denote iso/hypoentropic cases. Results of Over-Representation Analysis of DM-DEGs are shown in panel B: the analysis stratifies genes by DMR genomic location (5'Reg, GB, 3'UTR) and CGI status. The correlation plot categorizes DMRs based on combinatorial shifts in methylation and entropy (e.g., hypermethylated/hyperentropic). Color intensity represents statistical significance (according to Fisher's exact test) relative to background gene sets (DMGs, DEGs, and total RNA-seq targets). In panel C we show results of ORA specifically for underexpressed DM-DEGs, following the stratification in panel B. Cells display the count of DM-DEGs, with color intensity indicating significance across cancer-related pathways (KEGG), transcription factors, tumor suppressors, and oncogenes (COSMIC, GEAR).

quantification of both β and entropy S .

Following methylation calling, we applied the BiSLM segmentation algorithm to detect DMRs between T0 and T48. As described in previous analyses, DMRs were categorized into the six classes previously described based on combined $\Delta\beta$ (hypermethylation or hypomethylation) and ΔS (hyperentropic, isoentropic or hypoentropic) variations.

BiSLM identified a total of 636 DMRs, of which the vast majority (616) were hypermethylated, while only 20 showed hypomethylation. Hypermethylated DMRs were also predominant in terms of cumulative genomic size (Figure 4.5A). When classified by entropy variation, most DMRs (approximately 80% of hypermethylated and 95% of hypomethylated) exhibited either isoentropic or hypoentropic profiles, indicating that epiallelic changes were predominantly associated with stable or reduced heterogeneity rather than increased disorder (Figure 4.5A). The overall size distribution of DMRs spanned from a few hundred base pairs to several kilobases, consistent with the resolution achieved in the AML dataset.

Annotation of DMRs using the PoreMeth2 functional module revealed that approximately half (52%) overlapped protein-coding genes (332 DMRs), while 8% were located within noncoding transcripts (including pseudogenes and lncRNAs) and 40% in intergenic regions. As observed in the AML analysis, the majority of DMRs localized in open chromatin and transcriptional regulatory sites, overlapping DNase hypersensitive sites (88%) and transcription factor binding sites (86%), whereas enhancer regions were involved in only a minor fraction (16%). Regarding CpG context, 34% of the DMRs overlapped CpG islands, while the remaining 66% were found in low CpG density regions (Figure 4.5A), further supporting the ability of long-read data to capture methylation variation in sparsely CpG-distributed genomic domains.

When focusing on DMGs, PoreMeth2's feature-priority annotation classified 305 genes across the three functional categories 5Reg, GB, 3UTR. The large majority of DMGs harbored hypermethylated DMRs located either within gene bodies (211 genes) or in 5 regulatory regions (86 genes), whereas only a minimal fraction mapped to 3UTRs (Figure 4.5B). Among 5Reg DMGs, most DMRs overlapped CpG islands (62 in CGIs vs 24 in non-CGIs), whereas hypermethylated DMRs within gene bodies primarily involved sparse CpG regions (131 in non-CGIs vs 80 in CGIs). Across all categories, the predominant epiallelic signature corresponded to hypoentropic

($\Delta S < -0.1$) or isoentropic ($-0.1 < \Delta S < 0.1$) DMRs, particularly within gene bodies located in low CpG density regions, where nearly 80% displayed a negative ΔS value (Figure 4.5C). These results indicate that the treatment-induced methylation changes were mostly accompanied by decreased epiallelic variability, consistent with selective stabilization of specific epiallelic configurations.

To evaluate the transcriptional consequences of these methylation alterations, we performed quadruplicate RNA-seq experiments on T0 and T48 samples using ONT sequencing and conducted differential expression analysis. Integration of transcriptomic and methylation data revealed only a small amount of DM-DEGs, and all of these were associated with hypermethylated DMRs. Interestingly, nearly all DM-DEGs exhibited hypoentropic or isoentropic methylation profiles ($\Delta S < 0.1$, Figure 4.5D), reinforcing the observation that transcriptionally relevant methylation events correspond to decreased or stabilized epiallelic diversity rather than random fluctuations.

We next examined the proportion of DM-DEGs relative to the total set of DEGs, stratifying them according to genomic context and CpG density. Significant enrichment of DM-DEGs was observed specifically for genes harboring DMRs in 5 regulatory regions overlapping CpG islands, and in gene bodies located within low CpG density regions (Figure 4.5E). This suggests that both promoter-associated methylation in CGIs and gene-body methylation in sparse CpG domains can exert measurable transcriptional effects when accompanied by coherent, entropy-reducing epiallelic changes. Conversely, DMRs with positive ΔS values did not correlate with transcriptional modulation, implying that they likely represent transient or non-functional methylation shifts.

Collectively, these findings confirm that the dual-metric approach implemented in PoreMeth2, integrating methylation frequency and entropy variation, enables the discrimination between biologically meaningful and stochastic methylation events. In the HPGC system, this capability allowed us to identify methylation changes associated with stable regulatory responses to GPCR stimulation, highlighting the generalizability of PoreMeth2 across both cancer-related and physiological models of epigenetic remodeling.

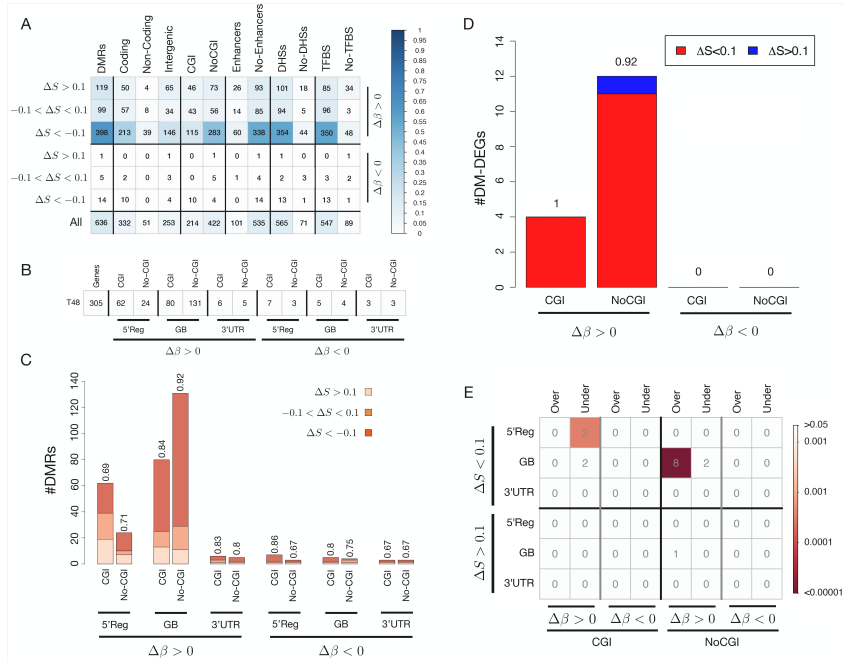


Figure 4.5: Characterization of DMRs in HPGC samples. The heatmap in panel A displays DMR counts across six categories defined by methylation ($\Delta\beta$) and entropy (ΔS) shifts. Rows represent overlaps with specific genic (coding, non-coding, intergenic) and regulatory (CGI, Enhancer, DHS, TFBS) elements for each category; color intensity indicates the proportional contribution of each category to the total DMR set. Panel B shows the count of DMGs associated with hypermethylated ($\Delta\beta > 0$) or hypomethylated ($\Delta\beta < 0$) DMRs, stratified by genomic feature (5'Reg, Gene Body, 3'UTR) and CpG island context. The stratification of the DMGs from Panel B into hyperentropic, isoentropic and hypoentropic classes is shown in panel C. Panel D represents the proportion of DM-DEGs exhibiting hyperentropic versus iso/hypoentropic patterns, grouped by methylation status and CGI overlap. Panel E shows the results of the Over-Representation Analysis of DM-DEGs, analyzed separately for genes affected by DMRs in 5' regulatory regions, gene bodies, and 3'UTRs, inside or outside CGIs. The plot displays results for four DMR classes defined by combinations of $\Delta\beta$ and ΔS . Cell numbers represent the total DM-DEG count per category, while color intensity reflects statistical significance.

4.4 Materials and Methods

4.4.1 AML samples: sequencing and data preparation

DNA from three pairs of matched acute myeloid leukemia (AML) samples, collected at diagnosis and relapse, was sequenced and processed as described in [97]. Basecalling and alignment were performed against the human reference genome (hg19), chosen for consistency with previous analyses and because the minor differences in CpG content between hg19 and newer assemblies do not substantially affect methylation-related analyses. 5mC were inferred using Nanopolish (v0.8.5) [91], applying log-likelihood ratio thresholds of < -2 and > 2 as recommended in the official documentation.

DMRs were identified using the PoreMeth2 pipeline, which integrates methylation frequency and entropy signals through the BiSLM algorithm. The resulting DMRs were further annotated according to genic and regulatory features as described in the main text.

For transcriptomic profiling, total RNA from the same AML pairs was sequenced on the Illumina NovaSeq 6000 platform as previously reported in [97]. Read quantification was performed using Salmon (v0.14.1) [108] against the GRCh37 Ensembl reference transcriptome. Normalization and differential expression analyses were conducted with DESeq2 (v1.30.1) [109]. Genes were considered differentially expressed if they exhibited an adjusted p -value < 0.05 (Benjamini-Hochberg correction) and an absolute $\log_2$ fold change > 0.5 .

4.4.2 HPGC samples: sequencing and data preparation

DNA extracted from HPGC cultures was sequenced before (T0) and after 48 hours (T48) of treatment with a GPCR agonist. Each library was sequenced on ONT r9.4.1 flow cells using a P2 Solo instrument with a 72-hour acquisition time per run.

Basecalling and modified-base calling were performed using Guppy (v6.5.7) with the super high accuracy (sup) model. Alignment to the human reference genome (hg19, for consistency with AML samples) was executed through the integrated minimap2 aligner (v2.24) [67]. Sex chromosomes were excluded from the analysis. Read-level 5mC likelihoods were extracted using Modkit (v0.4.1; <https://github.com/nanoporetech/modkit>), and methylation calls were filtered following the official guidelines provided in the Modkit

documentation.

For transcriptomic analysis, RNA was extracted from both T0 and T48 cultures (quadruplicate biological replicates) and sequenced on the same ONT platform. Basecalling and alignment were again performed with Guppy (v6.5.7), and read quantification was carried out using the `featureCounts` function from the Bioconductor `Rsubread` package (v2.12.3) [110]. Normalization and differential expression analysis were performed using DESeq2, with DEGs defined as those with adjusted p -value < 0.05 and absolute $\log_2$ fold change > 0.5 .

4.4.3 Over-Representation Analysis

ORA were performed to identify enrichment of DM-DEGs within known cancer-related pathways and gene sets. The analysis included pathways from the “Pathways in Cancer” collection of the KEGG database [107], and oncogenic signaling pathways from The Cancer Genome Atlas (TCGA). Gene lists were retrieved from the following sources:

- KEGG (<https://www.kegg.jp/kegg/download/>),
- COSMIC (https://cancer.sanger.ac.uk/cosmic/file_download),
- GEAR (<http://gear.comp-sysbio.org>),
- TFs from RegNetwork (<http://regnetworkweb.org/>). ORA was conducted using Fisher’s exact test, with the complete set of UCSC-annotated genes used as the background population.

4.5 Conclusion

In this chapter I reported PoreMeth2, the computational framework developed by our laboratory for the detection and interpretation of DMRs from long-read methylation data. The method is designed for pairs of evolutionarily related samples such as tumor versus normal tissues or pre- versus post-treatment conditions, where epigenetic alterations arise through progressive selection or remodeling rather than random fluctuation.

PoreMeth2 introduces a bivariate segmentation strategy that jointly models changes in methylation frequency ($\Delta\beta$) and methylation entropy (ΔS). While $\Delta\beta$ measures directional methylation shifts, ΔS captures alterations

in the epiallelic composition, distinguishing selective from stochastic processes. These two layers are integrated through the BiSLM algorithm, an HMM-based approach that segments the genome into coordinated regions of methylation and entropy change. Simulations showed that BiSLM detects DMRs with high accuracy at coverage $\geq 20\times$, resolving events involving as few as five CpGs even in sparse regions.

Implemented as an R package, PoreMeth2 also provides a multilayered annotation module intersecting DMRs with genic and regulatory elements, thereby supporting both statistical detection and functional interpretation.

We evaluated PoreMeth2 on two sets of evolutionarily related samples (AML and HPGC) and the results demonstrated that BiSLM is able to correctly identify DMRs also in low CpG density regions by integrating $\Delta\beta$ and ΔS , expanding resolution beyond short-read bisulfite approaches. Moreover, regions with $\Delta S < 0.1$ correlated most strongly with transcriptional modulation, indicating that entropy variation discriminates functionally relevant from random methylation changes.

Together, these results establish PoreMeth2 as a robust and biologically meaningful tool for long-read methylation analysis. By combining methylation frequency and entropy, it distinguishes selective epigenetic fixation from random methylation drift, offering deeper insight into the functional consequences of methylation changes.

4.5.1 Future developments

Future developments will focus on extending PoreMeth2 to broaden its applicability. In particular, we plan on introducing a new module to perform differential methylation analyses across groups of samples reflecting different experimental conditions (e.g., case versus control sets). In this context, methylation entropy will not be employed, as it is both computationally demanding and conceptually unsuitable for comparisons between unrelated samples, where the goal is not to capture evolutionary epigenetic remodeling but rather to identify stable marker regions. To date, the only available tool offering a similar functionality on Nanopore data are PycoMeth and Modkit. However, PycoMeth does not perform DMR discovery but evaluates $\Delta\beta$ values over user-defined genomic regions, and Modkit's segmentation algorithm is outperformed by PoreMeth2 even on sample pair analysis.

Additionally, we plan to include dedicated visualization modules to facilitate graphical exploration of DMR distributions, methylation–entropy rela-

tionships, and gene-level annotations, moving toward a comprehensive and user-friendly platform for integrative long-read epigenomic analysis.

Chapter 5

Epitranscriptomics

*Nanopore direct RNA sequencing offers the unique capability to detect post-transcriptional modifications on native molecules. However, the accurate decoding of the raw ionic current remains a significant computational challenge, currently constrained by the reliance on supervised models with high false discovery rates or the experimental burden of comparative case-control designs. In this chapter, we present a preliminary exploratory analysis focused on the stability of signal extraction strategies, a critical prerequisite for robust modification detection. By systematically comparing the signal segmentation provided by standard basecalling models against explicit Dynamic Time Warping re-alignment, we assess the reproducibility of the physical measurement. Our observations suggest that the basecaller’s intrinsic segmentation is highly model-dependent and insufficient for quantitative analysis, whereas explicit re-alignment appears necessary to ensure signal consistency. Furthermore, we provide initial evidence supporting the potential of read-guided (self) alignment strategies. These preliminary findings offer a methodological basis for future investigations into unsupervised, control-free frameworks capable of characterizing epitranscriptomic marks through the analysis of internal signal anomalies.*¹

¹This work was conducted while the author was a visiting PhD student at the German Cancer Research Center (DKFZ), Heidelberg (Germany), from March to June 2025, working with the Division of Immune Diversity (D150) lead by Prof. Nina Papavasiliou.

5.1 Introduction

5.1.1 RNA Editing and the Epitranscriptome

Historically, transcriptomic analyses were largely limited to the quantification of transcript abundance for gene expression profiling. However, the advent of high-sensitivity technologies has facilitated a shift towards the multidimensional characterization of RNA: current methodologies allow for the simultaneous detection of sequence, structure, and chemical modifications, thereby capturing the functional heterogeneity of the transcriptome at the post-transcriptional level.

While the primary sequence of RNA is transcribed from DNA, its final functional state is determined by multiple tiers of post-transcriptional processing. Within this broader regulatory landscape, which includes alternative splicing and non-coding RNA-mediated regulation, the epitranscriptome represents a distinct layer of chemical complexity. This specific regulatory level encompasses both RNA editing, which alters the nucleotide sequence, and reversible chemical modifications, which decorate RNA bases without changing the underlying code. Together, these alterations fine-tune RNA structure and interaction potential, thereby modulating the fate and function of both coding and non-coding transcripts in response to cellular cues.

RNA Editing

RNA editing represents a unique form of post-transcriptional processing where the nucleotide identity of a transcript is changed after transcription. In mammals, this phenomenon is primarily orchestrated by two enzymatic families: the AID/APOBEC proteins, which catalyze cytidine-to-uridine (C-to-U) deamination, and the ADAR proteins, responsible for adenosine-to-inosine (A-to-I) conversion. A-to-I editing is the most prevalent modification in the human transcriptome [111], [112], while C-to-U editing is essential for fewer specific physiological functions (most notably in regulating lipid metabolism through the site-specific Apolipoprotein B mRNA truncation [113]).

A-to-I editing involves the hydrolytic deamination of adenosine, preferentially on double-stranded RNA regions (dsRNA). Crucially, the cellular machinery, including the translation apparatus and the spliceosome, interprets inosine as guanosine (G) due to their similar base-pairing properties

with cytidine. As a consequence, this conversion functions effectively as a post-transcriptional A-to-G mutation [114].

The function and biological impact of this recoding varies profoundly depending on the genomic context. In protein-coding regions, particularly within the central nervous system, A-to-I editing can drive site-specific amino acid substitutions and therefore alter protein function. For instance, A-to-I editing of the *GRIA2* transcript (which encodes a subunit of the AMPA glutamate receptor) causes a glutamine to arginine substitution in the translated protein. This single modification is essential for the regulation of synaptic plasticity and neurotransmission in neurons [115].

However, the vast majority of editing events occur outside coding regions, specifically within Alu repetitive elements that tend to form long dsRNA structures. In this context, ADAR1 serves a critical immunoprotective role: by introducing mismatches into these dsRNA molecules, editing destabilizes the structure and prevents it from being recognized as viral genetic material by cytosolic immune sensors like MDA5. This self-masking mechanism is essential for maintaining immune homeostasis and the failure of this process triggers aberrant interferon responses, linking ADAR dysfunction to severe autoinflammatory disorders, neurodegeneration and cancer [116]–[118].

Chemical modifications of RNA

In parallel with RNA editing, RNA molecules are subject to over 170 discrete chemical modifications that do not alter the sequence but profoundly influence RNA structure, stability, and function [119]. These modifications are prevalent across all types of RNA, but one of the most extensively studied in messenger RNA is N6-methyladenosine (m6A), which affects multiple aspects of RNA metabolism, including splicing, translation, localization, and decay [120], [121].

Other modifications (such as 5-methylcytosine (m5C), pseudouridine (Ψ), and 2'-O-methylation (Nm)) contribute to the structural stability of RNA and regulate translation and immune responses [122]. Although the roles of many of these marks are still being elucidated, it is increasingly clear that the epitranscriptome represents a key layer of post-transcriptional gene regulation.

RNA editing and RNA modifications are now recognized as integral components of the post-transcriptional regulatory landscape, expanding the functional capacity of the transcriptome beyond genomic encoding. Editing can

influence methylation patterns and vice versa, as recent studies suggest a crosstalk between these processes that impacts RNA secondary structure, processing, and fate [123], [124]. Together, these mechanisms leverage the plasticity of RNA to enable dynamic and context-specific gene regulation.

5.1.2 Evolution of Detection Methodologies

Biochemical Assays and NGS

Given their pivotal role in regulating cellular functions, the detection of RNA modifications has been a core objective of molecular biology for decades. Consequently, the technological landscape in this field has undergone a dramatic transformation, evolving from low-throughput biochemical assays to high-throughput transcriptomic mapping.

Historically, detection relied on low-throughput biochemical assays (such as Thin Layer Chromatography and Mass Spectrometry) or laborious site-specific protocols like SCARLET [125], [126]. While chemically precise, these methods lacked the scalability required for transcriptome-wide mapping, effectively limiting the field to the study of isolated targets rather than global regulatory networks.

The advent of NGS represented a first revolution in the field of epitranscriptomics, enabling the first transcriptome-wide maps of RNA modifications. However, because standard reverse transcriptases cannot distinguish most modified bases from canonical ones, NGS-based approaches also rely on indirect detection strategies that convert the modification status into a readable signal before sequencing.

A pioneering strategy involved the use of specific antibodies to immunoprecipitate RNA fragments containing the modification of interest, a technique best exemplified by MeRIP-Seq (or m6A-seq) [127]. While this approach effectively allowed to generate the first global modification maps, it is inherently limited by low spatial resolution, identifying broad 100-200 nucleotide regions (or peaks) rather than single bases, precluding the precise determination of modification stoichiometry at specific sites.. Moreover, these methods are constrained by high input material requirements and potential antibody cross-reactivity.

To achieve single-nucleotide resolution, strategies based on chemical conversion (originally established for DNA analysis) have been adapted to transcriptome analyses. These methods leverage specific chemical reactions to

alter the base-pairing properties of modified nucleotides, translating epigenetic marks into readable sequence changes or reverse transcription stops.

Prominent examples include bisulfite sequencing applied to RNA (BS-RNA-seq), which allows for the precise mapping of 5-methylcytosine (m5C) [128], [129], and carbodiimide (CMC) derivatization, which is the gold standard for detecting Pseudouridine (Ψ) [130], [131]. Similarly, Inosine Chemical Erasing (ICE-seq) exploits acrylonitrile treatment to selectively cyanoethylate Inosine residues, distinguishing genuine A-to-I editing events from sequencing errors [132].

However, while these techniques overcome the resolution limits of antibody-based assays, they rely on harsh chemical treatments that are often incompatible with the labile nature of RNA. For instance, the high temperatures and acidic pH required for efficient bisulfite conversion, while tolerable for DNA, cause severe hydrolysis of the phosphodiester bonds in RNA. This degradation leads to fragmented libraries, loss of complexity, and a bias towards highly abundant or shorter transcripts, ultimately complicating the quantification of modification stoichiometry.

Crucially, both approaches share a fundamental bottleneck: the reliance on cDNA conversion and PCR amplification, which eliminate modifications on the native molecule. Therefore, an intrinsic limitation of NGS-based approaches is the loss of information about combinatorial patterns (phasing) of modifications on the same strand and the introduction of amplification biases.

The Paradigm Shift: Direct Native RNA Sequencing

As previously described in Chapter 1, ONT currently provides the only commercially available platform for the direct sequencing of native RNA molecules (dRNA) [133]. By bypassing the reverse transcription and amplification steps required by short-read technologies, this approach preserves the chemical structure of ribonucleotides during data acquisition [134]. Consequently, the ionic current signal inherently contains information regarding both the primary nucleotide sequence and the presence of post-transcriptional modifications, enabling, in principle, the study of modification stoichiometry and phasing on single molecules [135].

However, as the application of Nanopore sensing to native RNA is a relatively recent development, its analytical reliability has not yet reached the maturity of DNA sequencing standards.

A primary limitation lies in the signal fidelity of the current dRNA chemistry. The translocation of RNA strands through the motor protein and the pore exhibits higher velocity and irregularity compared to DNA, resulting in reduced signal-to-noise ratios [136]. Consequently, the raw read accuracy of dRNA typically ranges between 90% and 93%, which is significantly lower than the >99% accuracy routinely achievable with modern DNA chemistries (e.g., Q20+) [137]. This error profile, characterized by a high frequency of insertions and deletions (indels), raises a challenge in the alignment of reads to reference genomes. Since modification calling algorithms rely on the precise registration of raw signal events to genomic coordinates (a process known as signal-to-reference alignment or "resquiggling"), mapping errors can lead to signal misalignment and the subsequent generation of artifacts in modification detection [138].

Furthermore, the diversity of the epitranscriptome presents a significant hurdle for signal decoding. While DNA methylation analysis is primarily restricted to 5-methylcytosine, RNA molecules harbor over 170 distinct chemical modifications [119]. Many of these modifications share similar chemical structures (e.g., stereoisomers) or induce subtle signal shifts that are indistinguishable from canonical bases or random signal fluctuations using current sensor technology [139]. Additionally, the propensity of RNA to fold into stable secondary structures can induce stalling or variable translocation speeds during sequencing; these dwell-time anomalies can be confounded with modification-induced signal delays, further complicating the classification tasks [140].

Finally, the development of supervised machine learning models for modification calling is constrained by the scarcity of "ground truth" datasets. Unlike DNA, where PCR amplification can generate strictly unmodified controls, producing reference RNA datasets with known modification patterns is methodologically challenging [141]. While In Vitro Transcription (IVT) can generate unmodified transcripts, the synthesis of long RNA molecules containing specific modifications at 100% stoichiometry (necessary for training neural networks) remains difficult for most modification types [142]. This lack of labeled training data forces many computational tools to rely on comparative strategies (e.g., detecting signal differences between IVT and native samples) rather than direct classification, thereby limiting their ability to identify modifications in isolation or within complex mixtures.

5.2 State of the Art

While the physical principle of Nanopore sequencing theoretically allows for the sensing of any RNA modification, the computational decoding of these signals represents a significant bioinformatics challenge. The raw output of the sequencer is a continuous measurement of ionic current (squiggle) sampled at high frequency. When a modified base passes through the pore, it induces a signal deviation compared to canonical nucleotides. However, this deviation is highly complex, dependent not only on the modification itself but also on the surrounding sequence context (k-mer).

To navigate this complexity, the computational landscape has diversified into three primary methodological categories:

- **Supervised Classification Models**

This category encompasses methods that rely on machine learning algorithms trained on labeled datasets to recognize modification signatures. The current industry standard, promoted by ONT, employs Deep Learning models integrated directly into the basecalling process. The state-of-the-art implementation is Dorado, which can output probability scores for modifications alongside the base sequence (often via the post-hoc module Remora). These "black box" algorithms offer high usability and speed but are strictly bound to the quality and diversity of their training data, often suffering from high False Discovery Rates (FDR) in underrepresented sequence contexts [143], [144].

A distinct approach within this category is represented by EpiNano [145], [146]. Unlike the end-to-end deep learning models of basecallers, EpiNano adopts a feature-engineering strategy rooted in the observation that chemical modifications induce systematic basecalling errors. By extracting specific metrics from the aligned reads (such as mismatch frequency, insertion/deletion rates, and current intensity), it employs a classical Support Vector Machine (SVM) to distinguish modified from unmodified bases. While this approach offers greater interpretability than deep neural networks, it remains dependent on the availability of high-quality training sets to learn the specific error profiles associated with each modification type.

- **Comparative Analysis (Sample vs. Control)**

To mitigate the biases of pre-trained models, the field has widely adopted Comparative Analysis as a robust standard for validation. These methods do not attempt to classify the modification directly but instead detect statistical differences in the raw signal features between a target native sample and an unmodified control (IVT).

Approaches in this category include tools such as Nanocompore [141], xPore [137], and Eligos [142]. These pipelines typically align the signal to the reference and perform statistical tests (e.g., Kolmogorov-Smirnov, G-test, or Student’s t-test) to identify genomic positions where the current intensity or dwell time distribution of the native sample diverges significantly from the control.

• Unsupervised and Model-Free Approaches

More recently, a third category has emerged aiming to bypass both the rigidity of supervised labels and the cost of control samples. These Model-Free approaches rely on anomaly detection algorithms (e.g., Isolation Forests, Autoencoders) to identify signal segments that deviate from the expected canonical distribution based solely on internal signal consistency [147]. By treating modifications as "anomalies" within the ionic current trace, these methods offer a theoretical framework for identifying novel or rare modifications without prior training. However, their efficacy is contingent upon the precise isolation of the raw signal segments corresponding to each nucleotide, a task that remains computationally non-trivial.

It is worth noting that the practical application of these methods is often hindered by complex installation and execution requirements. Tools like MasterOfPores [148], [149] have been developed to mitigate these barriers by providing a unified, Nextflow-based interface. This pipeline orchestrates the execution of established detection algorithms within a reproducible environment, handling the necessary preprocessing steps in a standardized and automated manner.

5.2.1 Resquigglng

While supervised models like Dorado perform an implicit alignment of the signal during the inference process, encoded in the *move* tag of the alignment file, this representation is inherently coarse-grained, as it results from

a strided and highly compressed mapping of the raw signal to the nucleotide sequence (as documented in the Dorado documentation [150]). As a consequence, move tables substantially reduce the temporal resolution and variability of the signal, making them insufficiently precise for downstream analyses that require fine-grained signal-to-sequence correspondence. Comparative and Model-Free approaches therefore require an explicit additional step of signal-to-reference alignment (also referred to as resquigglng). This process involves the precise extraction and mapping of the raw ionic current segments to specific genomic positions.

The raw output files (originally FAST5, now evolved into the more efficient POD5 format) contain the continuous ionic trace, but this data is not intrinsically mapped to the genomic sequence: to bridge this gap, the raw signal events must be realigned to the basecalled sequence. Methods developed to perform this operation include Tombo [92], which pioneered the resquigglng approach, and aligners like Nanopolish [91] (and its high-performance implementation f5c [151]) that employ a HMM-based algorithm called Adaptive Banded Event Alignment (ABEA).

While these established frameworks defined the standards for signal analysis derived from earlier RNA chemistries (RNA002), most of them have either been officially deprecated or currently lack the specific pore models required to process the complex signal topology of the latest RNA004 flow cells.

Consequently, the landscape of tools capable of performing this task on modern datasets is sparse. Currently, the only two methods providing pore models compatible with RNA004 chemistry are uncalled4 [152] and f5c [151], which rely on fundamentally different algorithmic strategies to solve the signal-to-reference alignment problem.

f5c employs the ABEA algorithm, originally developed for Nanopolish. ABEA operates within a HMM framework, where the reference sequence determines the emission probabilities (based on the expected current levels for each k-mer). The algorithm seeks the most likely path of events through the k-mer sequence. To ensure computational tractability, the search space is restricted to an adaptive "band" around the optimal alignment diagonal, dynamically adjusting the constraints based on the alignment score.

In contrast, uncalled4 is based on Dynamic Time Warping (DTW). Unlike the probabilistic HMM approach, DTW treats the alignment as a non-linear optimization problem, aiming to minimize the cumulative distance between

the observed raw signal trace and the synthetic trace predicted by the pore model. By locally compressing or expanding the time axis, DTW effectively handles the stochastic variation in polymer translocation speed (dwell time) without requiring the strict state transitions of an HMM.

This alignment precision comes at a steep computational cost, which is particularly relevant considering that detection methods are based on either comparative experimental designs (requiring negative control samples) or model-free approaches that rely on constructing a large "corpus of truth" from unmodified samples.

In light of these challenges, the results presented in this chapter focus on a comparative analysis of signal extraction strategies. Specifically, we evaluate the consistency and fidelity of raw signals extracted directly from basecaller move tables (using a custom implementation) against those obtained via explicit realignment algorithms (Uncalled4 and f5c).

5.3 Results

5.3.1 Signal Extraction and Realignment Assessment

The primary objective of this study was to evaluate the fidelity and stability of different signal extraction pipelines for dRNA sequencing data. Accurate signal segmentation is the foundational step for any downstream epitranscriptomic inference, particularly for model-free detection methods that rely on deviations from a baseline corpus of truth. To isolate technical variability from biological noise such as naturally occurring post-transcriptional modifications, we employed an IVT RNA dataset. Since IVT RNA is synthesized using canonical nucleotides, any observed variance in the signal distribution can be attributed to the sequencing chemistry, the basecalling model, or the signal extraction algorithm itself.

We designed a comparative framework to evaluate two state-of-the-art realignment tools, Uncalled4 and f5c, against the native output provided by the Dorado basecaller. The analysis was conducted using the Dorado basecaller with the three available models with varying accuracy: fast, HAC and SUP. This multi-model approach allowed us to quantify the dependency of signal extraction quality on the underlying sequence accuracy.

We implemented three distinct extraction pipelines, by means of custom Python scripts to standardize the output formats:

1. Dorado Raw Alignment: Signals were extracted directly according to the `move` table generated by Dorado. This dataset represents the raw assignment of current samples to sequence bases as determined by the basecaller’s neural network, without any post-hoc realignment. We developed a custom parser to map these move indices to genomic coordinates based on the BAM alignment.
2. Uncalled4 (DTW): We processed the raw POD5 files using the `uncalled4 align` function, which employs DTW to align raw signal events to the reference sequence. The resulting alignment layers were parsed to extract the mean current and dwell time for every reference position.
3. f5c (ABEA): The dataset was processed using the `f5c eventalign` function, based on the ABEA algorithm. As for uncalled4 output, a dedicated script was written to parse the eventalign output and map event means to their corresponding genomic indices.

To evaluate biological coherence, we aggregated the extracted signals by genomic position. For each nucleotide in the reference genome, we calculated the median ionic current and the standard deviation (SD) across all covering reads. These metrics serve as proxies for the stability of the signal level and the precision of the alignment, respectively.

The visual inspection of the signal traces (Figure 5.1) on an example genomic region reveals a significant divergence between the raw basecaller output and the realignment tools, with the Dorado raw alignment exhibiting high-frequency stochastic fluctuations. While this preserves the instantaneous nature of the sensor readings, it introduces substantial noise that complicates the definition of a stable baseline. In contrast, both Uncalled4 and f5c produce congruent smoothed median current profiles, indicating that while realignment is a critical step for filtering technical noise and standardizing signal distributions, DTW and ABEA provide globally comparable results.

Quantitatively, we assessed the robustness of these methods by calculating the correlation of signal features extracted using the Fast model versus the SUP model, to determine the impact of basecalling accuracy on each pipeline (Figure 5.2). Our analysis indicates that reference-guided realignment tools maintain high correlation ($R \sim 0.95$) among models across genomic positions, implying that the realignment step effectively mitigates

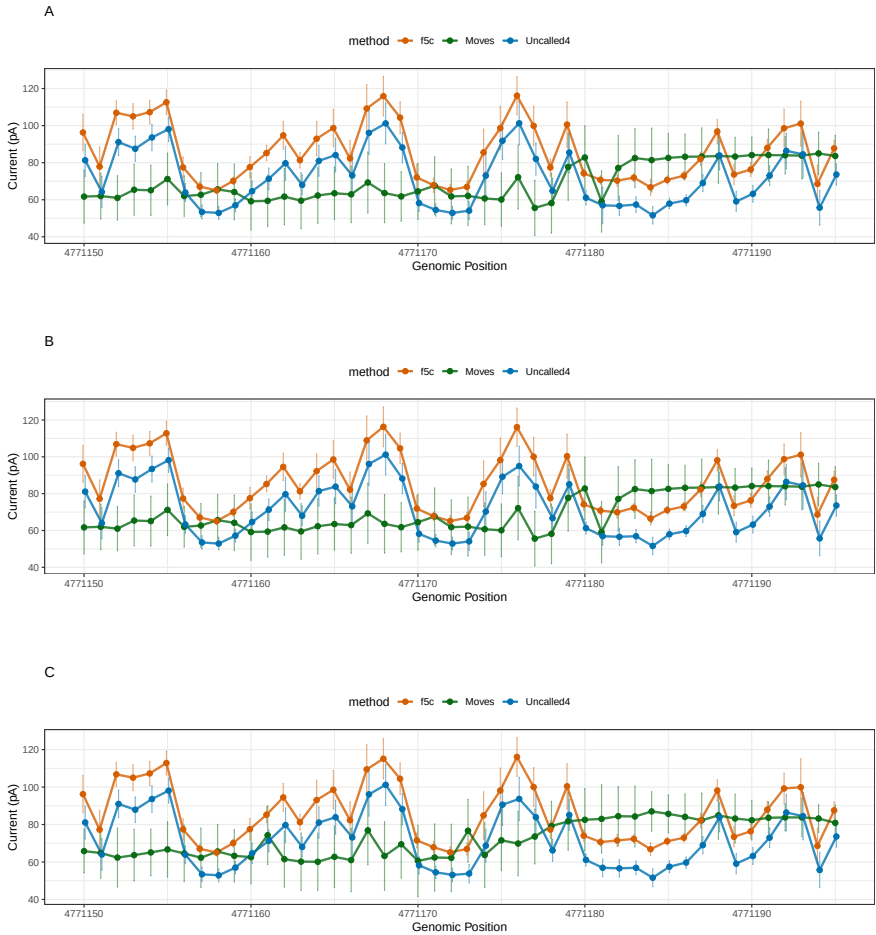


Figure 5.1: Comparison of median ionic current profiles across extraction frameworks on an example genomic region. Results are shown for the Fast (A), HAC (B) and SUP (C) basecalling models. Realignment tools (Uncalled4 and f5c) demonstrate significantly higher signal coherence compared to the raw basecaller output and show comparable profiles among basecalling models.

inaccuracies introduced during basecalling by anchoring the signal to the expected reference k-mer rather than to the basecalled sequence. This suggests that both uncalled4 and f5c are largely robust to basecalling errors when operated in a reference-guided setting. Conversely, raw extraction showed a significantly lower correlation ($R \sim 0.55$), highlighting its susceptibility to basecalling errors.

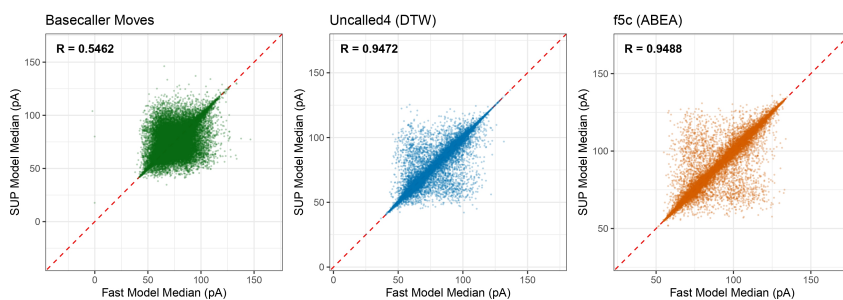


Figure 5.2: The figure shows the correlation between signals extracted from reads basecalled with the Fast (least accurate) and SUP (most accurate) models for raw aligned data (extracted via the `move` table), uncalled4 and f5c.

To rigorously validate the reproducibility of the extracted signals, we performed an internal consistency test. Genomic positions sharing the same reference k-mer were randomly partitioned into two independent subsets (Subset A and Subset B), and the mean signal was computed for each genomic position. We then assessed the correlation between the aggregated signals of the two subsets. This correlation reflects the extent to which identical k-mers exhibit consistent signal profiles across distinct genomic contexts, thereby providing a quantitative measure of the reproducibility of the extraction method and of the influence of broader sequence context on k-mer-associated signals.

As shown in Figure 5.3, data extracted according to the Dorado `move` table shows a relatively low correlation between subsets ($R \sim 0.68$), confirming that raw Dorado alignment is too noisy for generating a reliable consensus without secondary processing. Conversely, both Uncalled4 and f5c exhibited robust internal consistency ($R \sim 0.96$).

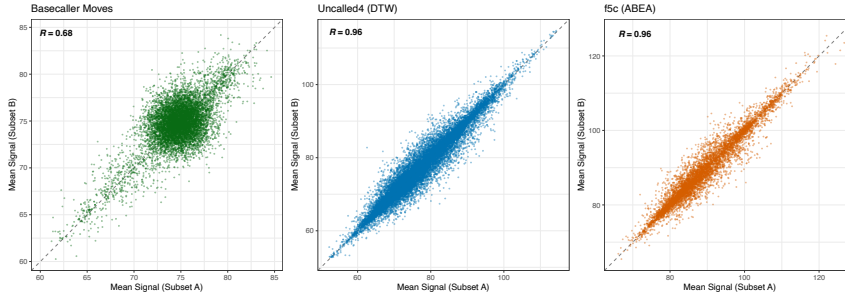


Figure 5.3: The figure shows the Internal consistency assessment of random subsets of genomic positions exhibiting identical k-mer sequence. The high correlation observed in realigned data validates the biological significance of signal extracted after resquigging, also confirming the similar performance obtained by uncalled4 and f5c. Conversely, as expected, raw signal alignment provides biologically less coherent values, resulting in a lower correlation between subsets.

5.3.2 Self-Alignment

Standard reference-guided alignment is based on forcing the raw signal to match the expected sequence provided by a reference genome. While this approach standardizes the data coordinates and allows to correct basecalling inaccuracies, it may introduce errors due to the reference bias. If the actual RNA molecule contains a variant or a biologically edited base that differs from the reference, the alignment algorithm may force-fit the signal segments to the wrong k-mer model, resulting in distorted current distributions.

To mitigate this bias, both f5c and uncalled4 offer an alternative Self-Alignment mode: in this mode, the signal is aligned not to the external genome, but to the read's own basecalled sequence. This read-based approach hypothesizes that the basecalled sequence, although potentially containing errors, is a more faithful representation of the physical molecule passing through the pore than a theoretical reference.

To test this strategy we developed an custom Python pipeline. Since Self-Alignment produces data indexed by the read's internal coordinates rather than genomic positions, we created a mapping framework that:

1. Extracts the signal events aligned to the read sequence (Self-Alignment).
2. Projects the signal events from the read-space to the genome-space

(according to read alignment).

3. Calculates signal median and standard deviation from all the events mapped to each reference position.

It is crucial to note that although the final metrics are reported per genomic position, the underlying signal segmentation is fundamentally different from the Reference-Guided approach. In Reference-Guided mode, the alignment algorithm is constrained by the reference k-mers; in Self-Alignment, the segmentation follows the raw basecalls.

As a first step, we evaluated the impact of the basecalling model accuracy on the extracted. As expected, self-realignment shows a higher dependency on basecalling model, with a lower global correlation ($R \sim 0.54$ for both uncalled4 and f5c) of median signal between fast and SUP basecalling compared to Reference-Guided realignment (see Figure 5.4).

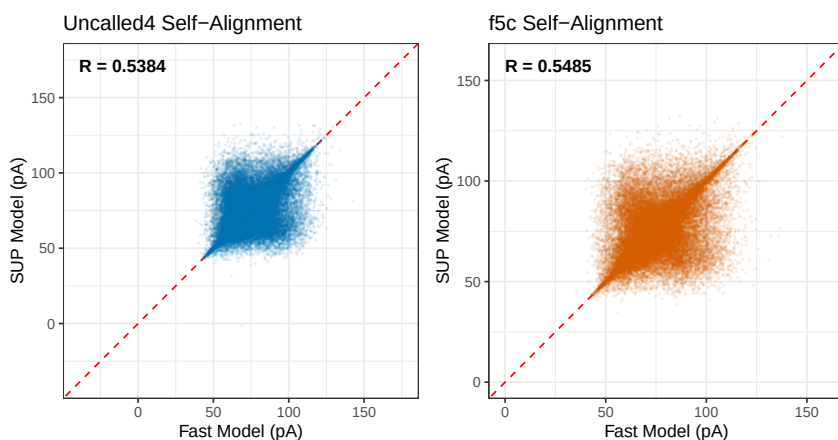


Figure 5.4: The figure shows the correlation between signals extracted in Self-Alignment mode from reads basecalled with the Fast (least accurate) and SUP (most accurate) models for uncalled4 and f5c. The lower correlation (compared to the one observed for both methods with the standard Reference-Realignment mode) confirms the higher dependency of Self-Realignment on sequence accuracy, and the lower ability to bypass basecalling errors.

Despite this increased dependency on basecalling accuracy, self-aligned signals retain high internal consistency when evaluated through subset correlation analyses ($R \sim 0.9$, Figure 5.5) both for uncalled4 and f5c, also

for the Fast (less accurate) basecalling model. This results suggest that self-alignment represents a viable strategy for reference construction when minimizing reference-induced bias is a priority.

5.3.3 K-mer based aggregation

A major bottleneck in developing model-free epitranscriptomic methods is the requirement for extremely high coverage at every genomic position to build statistically significant signal distributions. Constructing a Corpus of Truth solely based on genomic coordinates is often inefficient, as many regions of the transcriptome are naturally low-expressed, leading to sparse data and unreliable baselines for detecting modifications.

To overcome this limitation, we proposed and evaluated an alternative aggregation strategy: pooling signal events based on k-mer identity rather than their unique genomic position. The underlying hypothesis is that the physicochemical properties of a k-mer (e.g., the ionic current blockade level) are largely determined by its local sequence context, regardless of its specific location in the genome. By aggregating all instances of a specific k-mer across the entire transcriptome, we aim to artificially increase the effective coverage by orders of magnitude, thereby generating robust signal distributions even from standard sequencing depths.

To validate this strategy, we processed the Self-Aligned dataset generated using Uncalled4 (as described in the previous section), grouping signal segments by their 9-mer sequence identity instead of their genomic coordinate. Global statistics, including median current and standard deviation, were then calculated for each of the observed 9-mers to construct a global k-mer reference. For simplicity and consistency, this analysis focused on the output from Uncalled4, which demonstrated high reliability in previous self-alignment tests.

As a first metric of success, we quantified the increase in data density achieved by aggregating on k-mer sequence. In our IVT dataset, the average positional coverage was approximately 5 reads per genomic site. In sharp contrast, the k-mer-based aggregation yielded a mean effective coverage of approximately 83,000 observations per k-mer. Here, an "observation" is defined as the product of the number of reads covering a given position and the number of genomic loci where that specific k-mer occurs. This substantial increase in coverage enables the construction of robust baseline signal distributions from standard whole-transcriptome sequencing data.

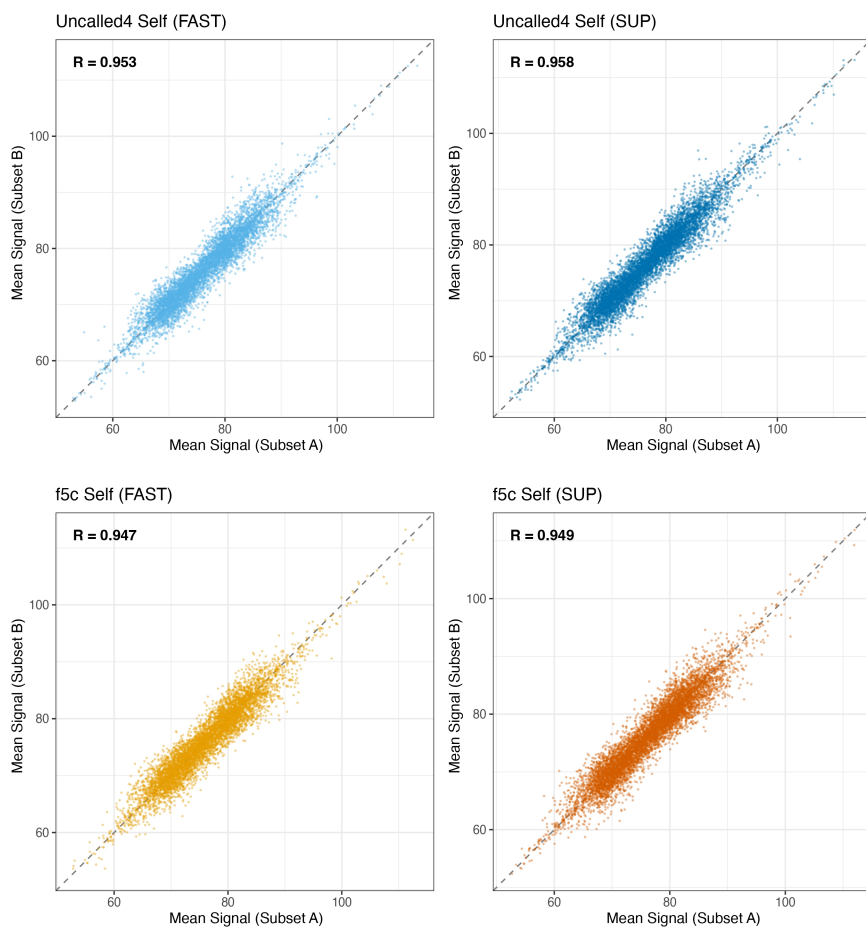


Figure 5.5: The figure shows the Internal consistency assessment of random subsets of genomic positions exhibiting identical k-mer sequence after Self-Alignment with uncalled4 and f5c on reads basecalled with the Fast and SUP models.

We next evaluated whether pooling signals from diverse genomic contexts introduced excessive noise, which would undermine the precision of the reference model. We compared the distribution of signal standard deviations for events aggregated by k-mer identity against those aggregated by single genomic positions.

As shown in Figure 5.6, the signal distributions for global k-mers exhibit a mean standard deviation comparable to that of site-specific signals (ranging from ~ 7.7 to ~ 8.4 pA). The overlapping density curves suggest that the local k-mer sequence is indeed the primary determinant of the ionic current, and that distal genomic context exerts a negligible effect on signal variance. This finding supports the use of a global k-mer model as a stable Corpus of Truth.

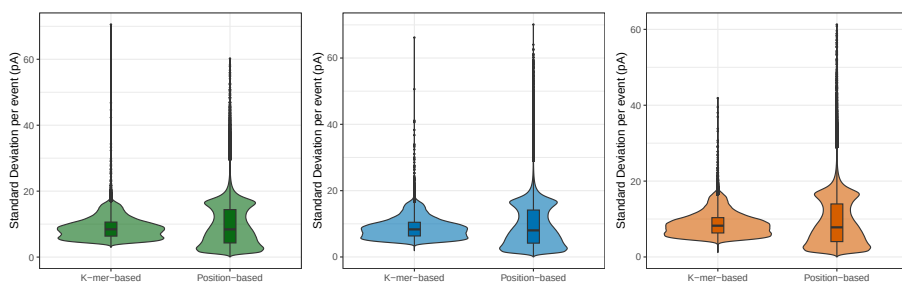


Figure 5.6: The figure shows the comparison of standard deviation from the median signal of each event (k-mer or genomic positions) for k-mer based and position-based signal aggregation. Results are shown for Fast (green), HAC (blue) and SUP (orange) basecalling models.

To rigorously validate the biological accuracy of the global k-mer model, we assessed the correlation between the signal values derived from this new aggregation method and the "ground truth" values obtained from position-specific genomic aggregation.

First, we verified the technical stability of the method by repeating the same comparative analysis performed in previous sections: correlating the global k-mer medians obtained from reads basecalled with the Fast model against those from the SUP model. This analysis yielded an extremely high correlation ($R = 0.995$, Figure 5.7), significantly surpassing the values observed in site-specific analysis. This high stability confirms that the k-mer aggregation strategy is robust to basecalling errors and produces highly con-

sistent signal estimates across different accuracy models, a result likely driven by the massive increase in observations per event which effectively neutralizes stochastic noise.

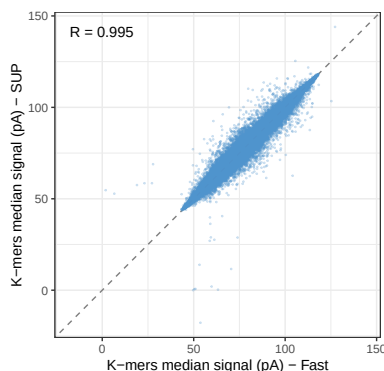


Figure 5.7: The figure shows the correlation of median signals after k-mer based signal aggregation performed on reads basecalled with the Fast and SUP basecalling models. The extremely high correlation confirms the stabilization of the signal when substantially increasing the number of observations.

Lastly, we calculated the correlation between the median signal assigned to each global k-mer and the median signal of genomic positions characterized by the same k-mer sequence. This analysis revealed a high degree of biological coherence, with Pearson correlation coefficients of $R = 0.893$ for the Fast model and $R = 0.909$ for the SUP model (Figure 5.8). This strong linear relationship further indicates that the average signal of a k-mer collected across the transcriptome is a statistically accurate predictor of the signal at any specific genomic locus, thereby legitimizing the use of a transcriptome-wide corpus of truth for detecting local anomalies.

5.4 Materials and Methods

5.4.1 Data Preparation and Processing

The dataset analyzed in this study consists of direct mRNA sequencing reads derived from murine macrophages. Specifically, we utilized the In Vitro Transcribed (IVT) dataset generated by Lemercier et al. [147]. The choice

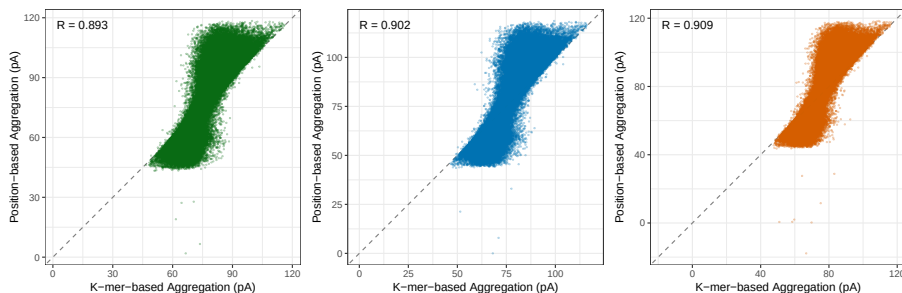


Figure 5.8: The figure shows the correlation between the median signal of each k-mer and the median signal of genomic positions exhibiting by the same k-mer sequence. Results are shown Fast (green), HAC (blue) and SUP (orange) basecalling models.

of IVT RNA was crucial to establish a modification-free baseline, allowing us to isolate signal variability originating solely from sequence context and technical factors.

Raw sequencing data (POD5 files) were processed using the ONT basecaller Dorado (v.0.8.1). In order to evaluate the impact of basecalling model complexity on signal segmentation accuracy, we performed basecalling using three distinct models provided by ONT: Fast, HAC and SUP. All basecalling runs were executed with the `--emit-moves` flag to preserve the move table within the output BAM files, enabling the downstream mapping of raw signal strides to basecalled nucleotides.

Prior to signal extraction, the alignment files were filtered to ensure high-quality data. We retained only primary alignments with a Mapping Quality score (MAPQ) greater than 10.

5.4.2 Signal Extraction and Realignment Pipelines

To compare different signal processing strategies, we implemented three distinct extraction pipelines. Custom Python scripts were developed to parse the outputs of each tool and standardize the signal metrics (median current, standard deviation) into a unified format for statistical analysis.

Signal-to-Reference realignment was performed using two tools:

- **Uncalled4:** The `uncalled4 align` (version 4.1.0) subcommand was executed with default parameters in `--eventalign-out` mode. This

generated an eventalign-like TSV file containing the mapping between reference k-mers and signal events. We then employed a custom Python script to parse this output, converting the standardized current levels back to raw picoampere (pA) values using the read-specific scaling parameters (scale, shift) extracted from the POD5 files.

- **f5c:** The `f5c eventalign` (version 1.6) subcommand was used with default options, adding the `--print-read-names` `--samples` flags to obtain calibrated signal values for each read position. Prior to execution, POD5 files were converted to the BLOW5 format using `pod5 convert to_fast5` and `slow5tools f2s`, and sequencing reads were indexed using `f5c index`. A dedicated parser was written to map the event means to genomic indices.

Signal-to-Read realignment (Self-Alignment) was performed using `uncalled4` and `f5c` using the following settings:

- **Uncalled4 Self-Alignment:** The `uncalled4 align` subcommand was used in `--self` mode. This produced an alignment of signal events to the read's k-mers. Our custom post-processing script then matched these "read k-mers" to the corresponding reference genomic positions based on the read's alignment (BAM), allowing us to compare self-aligned signals in a genomic context.
- **f5c Self-Alignment:** The `f5c resquiggle` command was used to align signals to the basecalled sequence. Similar to the Uncalled4 pipeline, we extracted the signal indices per read position and mapped the corresponding calibrated signal values to reference positions using the alignment information.

5.4.3 Custom Signal Extraction Pipelines

To standardize the signal data obtained from different sources and realignment strategies, we developed a group of custom Python scripts. These pipelines were designed to parse the raw outputs from Dorado, Uncalled4, and f5c, and to perform two distinct types of data aggregation: by genomic position and by k-mer sequence.

For the raw extraction method, signals were segmented directly according to the move table generated by Dorado during basecalling. We developed a custom Python code to parse the `move` table along with additional relevant

tags (ts, ns) and maps them to the raw signal array in the corresponding POD5 file. Raw signal segment are then assigned to each base called by the network and mapped to genomic coordinates based on the BAM alignment CIGAR string. This pipeline represents the "native" signal assignment as determined by the basecaller's neural network, without any post-hoc refinement.

For Uncalled4 and f5c outputs, which provide signal data in standardized TSV formats (eventalign or similar), our scripts parse the event tables to extract the mean current and standard deviation for each event (k-mer or genomic position). Raw current values are calibrated using the read-specific scaling parameters (scale, shift) extracted from the POD5 files to ensure comparability across reads.

Aggregation by Genomic Position

For position-based analyses, extracted signal events are mapped to reference coordinates using the read alignment information (BAM/CIGAR string). Our pipeline groups all signal events covering a specific genomic coordinate and calculates the median current and standard deviation across all covering reads. This approach constructs a site-specific signal profile.

Transcriptome-wide K-mer Aggregation

To implement the k-mer-based strategy, we developed a specialized pipeline that aggregates signals based on sequence identity rather than position. This pipeline processes the output of Self-Aligned reads and groups signal events based on their 9-mer sequence (observed on the read), effectively decoupling the signal from its genomic coordinate. As a last step, the global median current, variance, and observation count for each of the unique k-mers found in the transcriptome is calculated.

5.4.4 Statistical Analysis and Performance Metrics

To quantitatively evaluate the stability, accuracy, and biological validity of the signal extraction methods, we employed the following statistical metrics.

To assess the robustness of signal extraction against basecalling errors, we performed a Pearson correlation analysis (R) between the signal profiles obtained from the Fast basecalling model and the SUP basecalling model. A

high correlation coefficient indicates that the extraction pipeline is consistent regardless of the basecalling model's complexity.

To assess the biological validity of the extracted signals, we developed an internal consistency metric based on k-mer identity. Genomic positions were grouped by their reference 9-mer sequence. For each unique k-mer with sufficient occurrences ($N > 50$), the associated genomic loci were randomly split into two subsets (Subset A and Subset B). The mean signal was calculated for each subset, and a linear regression was performed between Subset A and Subset B. The Pearson correlation coefficient of this regression serves as a proxy for signal precision.

To validate the accuracy of the k-mer aggregation strategy, we calculated the Pearson correlation between the global median signal of a k-mer (derived from the Transcriptome-wide aggregation) and the median signal of specific genomic positions containing that same k-mer. This metric assesses whether the global average is a reliable predictor of local signal behavior.

5.5 Conclusion

The comparative analysis presented in the previous chapter demonstrated that, in terms of raw signal reconstruction accuracy, uncalled4 (DTW) and f5c (ABEA) are functionally equivalent. Both algorithms converge on a nearly identical physical topology of the ionic current traces, validating the reliability of either approach for standard reference-based realignment. However, beyond pure signal fidelity, the choice between these two tools involves critical operational trade-offs regarding workflow integration, computational efficiency, and software maturity.

A major distinction lies in data handling and pipeline integration. uncalled4 is inherently designed for the modern Nanopore ecosystem: it natively supports the efficient POD5 raw data format and interacts directly with standard BAM alignment files. This allows for a streamlined pipeline where the rich metadata embedded in the BAM (such as methylation tags or movement tables) is preserved and utilized without redundancy.

In contrast, f5c remains anchored to legacy file structures. Currently, it does not support the direct processing of POD5 files, necessitating a mandatory conversion step to the older FAST5 format or the community-standard SLOW5/BLOW5 formats. This conversion introduces a significant overhead in terms of storage management and processing time. Furthermore, f5c re-

quires both the mapping file (BAM) and the primary sequence file (FASTQ) as inputs. In modern pipelines (e.g., Dorado), the FASTQ format is often bypassed in favor of uBAM (unaligned BAM) to prevent metadata loss, making f5c's requirement an architectural bottleneck that complicates the analysis workflow.

From a computational perspective, f5c maintains a clear advantage. Built as a high-performance adaptation of Nanopolish, it is highly optimized for parallel processing and, crucially, offers full GPU acceleration. This makes it the superior choice for large-scale population studies where throughput is the limiting factor. The HMM-based ABEA algorithm, confined to a narrow adaptive band, is computationally less expensive than the global optimization required by Dynamic Time Warping. Consequently, while uncalled4 is computationally tractable for transcriptomic datasets, it incurs a higher CPU cost per read, which may scale poorly for high-depth sequencing without dedicated optimization.

Furthermore, the "readiness" level of the software packages must be considered. f5c benefits from years of community testing and optimization, resulting in a highly stable codebase that is robust to edge cases. Conversely, uncalled4, being a more recent development aimed at bridging the gap for RNA004 chemistry, is currently in an active phase of evolution. While it introduces innovative features like Signal-to-Read (self) alignment, our analysis highlighted that these experimental modes currently exhibit variable stability. We encountered intermittent runtime exceptions and inconsistencies when processing complex alignment scenarios, which is characteristic of software that has not yet reached full maturity. While these limitations are expected to be resolved in future updates, they currently necessitate careful supervision of the pipeline, whereas f5c offers "production-grade" reliability for standard reference-based tasks.

Finally, our analysis suggests a promising alternative for the construction of baseline reference models. We demonstrated that aggregating signals by k-mer identity across the entire transcriptome yields a massive increase in effective coverage without sacrificing signal precision or biological accuracy. This strategy leverages the redundancy of k-mers in the transcriptome to generate statistically robust signal distributions even for rare motifs or low-abundance transcripts. By decoupling the signal model from specific genomic coordinates, this approach overcomes the coverage bottlenecks inherent in

standard RNA-seq, paving the way for more sensitive and comprehensive model-free detection of epitranscriptomic modifications.

Chapter 6

Conclusion

6.1 Summary of contribution

The work presented in this thesis explores the development of computational strategies for the analysis of long-read sequencing data, with a particular focus on structural variation and DNA methylation profiling. In recent years, third-generation sequencing technologies have deeply transformed genomics by enabling the direct interrogation of native DNA molecules at unprecedented resolution. Yet, their full potential still depends critically on the availability of analytical methods capable of overcoming intrinsic challenges such as elevated error rates, alignment uncertainty, and signal-level variability. Within this context, this thesis contributes novel methodological frameworks that enhance robustness, interpretability, and biological insight across distinct layers of genome regulation.

The first major contribution of this work is the development of GASOLINE, a computational method for the detection and characterization of structural variants from Oxford Nanopore sequencing data. By introducing the Normalized Reciprocal Overlap metric and leveraging a graph-theoretic clustering approach, GASOLINE addresses one of the persistent limitations of existing SV callers: the imprecision and redundancy of breakpoint predictions caused by the noise inherent to long-read alignments. Through extensive benchmarking on both synthetic and real datasets, GASOLINE demonstrated competitive or superior performance compared to state-of-the-art tools, with particular advantages in breakpoint accuracy, clustering stability, and detection of events at low coverage. A second key component of the

method is the implementation of a dedicated somatic SV module, which fills an important gap in current computational genomics pipelines. By explicitly comparing matched tumor-normal sample pairs through an NRO-based similarity measure and a statistical significance test, GASOLINE enables reliable detection of somatic rearrangements, even in heterogeneous and subclonal cancer genomes. Collectively, this work establishes GASOLINE as a versatile and accurate solution for both germline and somatic SV discovery, demonstrating how long-read data can be used not only to detect rearrangements but also to reconstruct their underlying haplotypic structure.

The second main contribution of the thesis concerns the detection and interpretation of differential DNA methylation from nanopore reads. Existing approaches for DMR calling are typically derived from bisulfite sequencing methodologies and rely almost exclusively on differences in average methylation levels. However, direct long-read methylation enables a far richer representation, capturing the epiallelic composition of each genomic region and revealing whether methylation changes occur uniformly across molecules or arise from clonal selection, epigenetic drift, or shifts in cell-type composition. To exploit this unique feature of long-read data, we developed PoreMeth2, a bivariate segmentation method that integrates changes in methylation frequency ($\Delta\beta$) with changes in epiallelic entropy (ΔS). This joint modeling strategy allows the discrimination between homogeneous and heterogeneous methylation alterations, which would remain indistinguishable under conventional univariate frameworks. Applications to both AML datasets and human peripheral glial cells confirmed that the combined analysis of $\Delta\beta$ and ΔS provides a more nuanced and biologically meaningful interpretation of epigenetic remodeling, highlighting cases where methylation changes reflect selective expansion of pre-existing epiallelic states rather than coordinated regulation. PoreMeth2 thus introduces a conceptual and methodological advance by reframing DMR detection in terms of the internal diversity of methylation patterns rather than solely their mean levels.

Finally, the scope of this research extended to a preliminary exploration of the epitranscriptomic landscape. Unlike the relatively well-established frameworks for DNA analysis, the direct detection of RNA modifications via Nanopore sequencing remains a developing frontier. Our work focused on exploring different computational strategies to address the complexities of signal segmentation in direct RNA sequencing. Our comparative assessment indicated that realignment tools, such as `uncalled4` and `f5c`, provide a stable

reconstruction of ionic current traces compared to raw basecaller outputs, effectively reducing technical noise. Building on this observation, we investigated the potential of a Transcriptome-wide K-mer Aggregation strategy as a means to overcome the coverage limitations inherent in standard RNA-seq data. Our preliminary analysis suggests that pooling signals based on sequence identity allows for a substantial increase in effective observation counts while maintaining signal coherence. This finding points towards the possibility of constructing an internal reference baseline directly from the experimental dataset, a methodological direction that, if further developed and validated, could help mitigate the dependency on external controls in model-free modification detection.

Taken together, the developments presented in this thesis demonstrate how long-read sequencing can serve as a unifying platform for structural, epigenetic, and epitranscriptomic genomics, and how novel computational methods are essential for unlocking the full richness of the signal. While the epitranscriptomic analysis remains exploratory, GASOLINE and PoreMeth2 contribute distinct, mature perspectives: the former reconstructs the large-scale structural changes that reshape genome architecture; the latter quantifies the fine-grained epigenetic dynamics that regulate gene activity and cellular identity. Both methods illustrate how moving beyond traditional short-read paradigms enables new biological questions to be addressed with higher resolution and interpretability.

Appendix A

Publications

This research activity has led to the following publications in international journals.

International Journals

- A. Magi et al., “Gasoline: Detecting germline and somatic structural variants from long-reads data.”, *Scientific Reports*, vol. 13, no. 1, p. 20817, 2023, doi: 10.1038/s41598-023-48285-0.
- G. Mattei et al., “Poremeth2 for decoding the evolution of methylome alterations with nanopore sequencing”, *Genome Research*, vol. 35, no. 11, pp. 2501-2512, 2025, doi: 10.1101/gr.280259.124. (Co-first author)
- M. Urtis et al., “Oxford Nanopore Technologies [ONT] Sequencing: Clinical Validation in Genetically Heterogeneous Disorders”, *Genes*, vol. 16, no. 11, p. 1325, 2025, doi: 10.3390/genes16111325.

Submitted

- I. Alfisi et al., “Benchmarking DNA Foundation Models for zero-shot variant effect prediction: the role of context, training, and architecture”, *bioRxiv*. (Accepted upon revision)

Bibliography

- [1] 1000 Genomes Project Consortium *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571, p. 68, 2015.
- [2] M. J. Chaisson, J. Huddleston, M. Y. Dennis, *et al.*, “Resolving the complexity of the human genome using single-molecule sequencing,” *Nature*, vol. 517, no. 7536, pp. 608–611, 2015.
- [3] P. H. Sudmant, S. Mallick, B. J. Nelson, *et al.*, “Global diversity, population stratification, and selection of human copy-number variation,” *Science*, vol. 349, no. 6253, aab3761, 2015.
- [4] R. Jaenisch and A. Bird, “Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals,” *Nature genetics*, vol. 33, no. 3, pp. 245–254, 2003.
- [5] S. Sharma, T. K. Kelly, and P. A. Jones, “Epigenetics in cancer,” *Carcinogenesis*, vol. 31, no. 1, pp. 27–36, 2010.
- [6] C. Ling and T. Rönn, “Epigenetics in human obesity and type 2 diabetes,” *Cell metabolism*, vol. 29, no. 5, pp. 1028–1044, 2019.
- [7] M. Nikolac Perkovic, A. Videtic Paska, M. Konjevod, *et al.*, “Epigenetics of alzheimer’s disease,” *Biomolecules*, vol. 11, no. 2, p. 195, 2021.
- [8] Q. Lu, “The critical importance of epigenetics in autoimmunity,” *Journal of autoimmunity*, vol. 41, pp. 1–5, 2013.
- [9] S. Heerboth, K. Lapinska, N. Snyder, M. Leary, S. Rollinson, and S. Sarkar, “Use of epigenetic drugs in disease: An overview,” *Genetics & epigenetics*, vol. 6, GEG–S12270, 2014.
- [10] A. Bird, “Dna methylation patterns and epigenetic memory,” *Genes & development*, vol. 16, no. 1, pp. 6–21, 2002.

- [11] L. D. Moore, T. Le, and G. Fan, "Dna methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, 2013.
- [12] F. Neri, S. Rapelli, A. Krepelova, *et al.*, "Intragenic dna methylation prevents spurious transcription initiation," *Nature*, vol. 543, no. 7643, pp. 72–77, 2017.
- [13] Q. Du, P.-L. Luu, C. Stirzaker, and S. J. Clark, "Methyl-cpg-binding domain proteins: Readers of the epigenome," *Epigenomics*, vol. 7, no. 6, pp. 1051–1073, 2015.
- [14] M. Okano, D. W. Bell, D. A. Haber, and E. Li, "Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development," *Cell*, vol. 99, no. 3, pp. 247–257, 1999.
- [15] N. Bhutani, D. M. Burns, and H. M. Blau, "Dna demethylation dynamics," *Cell*, vol. 146, no. 6, pp. 866–872, 2011.
- [16] M. Ehrlich and M. Lacey, "Dna hypomethylation and hemimethylation in cancer," *Epigenetic alterations in oncogenesis*, pp. 31–56, 2012.
- [17] V. Kloten, B. Becker, K. Winner, *et al.*, "Promoter hypermethylation of the tumor-suppressor genes itih5, dkk3, and rassfla as novel biomarkers for blood-based breast cancer screening," *Breast Cancer Research*, vol. 15, no. 1, R4, 2013.
- [18] Z. Jin and Y. Liu, "Dna methylation in human diseases," *Genes & diseases*, vol. 5, no. 1, pp. 1–8, 2018.
- [19] M. V. Greenberg and D. Bourc'his, "The diverse roles of dna methylation in mammalian development and disease," *Nature reviews Molecular cell biology*, vol. 20, no. 10, pp. 590–607, 2019.
- [20] R. Lister, M. Pelizzola, R. H. Downen, *et al.*, "Human dna methylomes at base resolution show widespread epigenomic differences," *nature*, vol. 462, no. 7271, pp. 315–322, 2009.
- [21] A. Kundaje, W. Meuleman, J. Ernst, *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, p. 317, 2015.
- [22] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, pp. 621–628, 2008.

- [23] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: A revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.
- [24] R. Stark, M. Grzelak, and J. Hadfield, "Rna sequencing: The teenage years," *Nature reviews genetics*, vol. 20, no. 11, pp. 631–656, 2019.
- [25] K. D. Meyer and S. R. Jaffrey, "The dynamic epitranscriptome: N 6-methyladenosine and gene expression control," *Nature reviews Molecular cell biology*, vol. 15, no. 5, pp. 313–326, 2014.
- [26] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.
- [27] A. M. Maxam and W. Gilbert, "A new method for sequencing dna.," *Proceedings of the National Academy of Sciences*, vol. 74, no. 2, pp. 560–564, 1977.
- [28] J. C. Venter, M. D. Adams, E. W. Myers, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.
- [29] J. Shendure and H. Ji, "Next-generation dna sequencing," *Nature biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.
- [30] E. R. Mardis, "Next-generation sequencing platforms," *Annual review of analytical chemistry*, vol. 6, no. 1, pp. 287–303, 2013.
- [31] M. Margulies, M. Egholm, W. E. Altman, *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.
- [32] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *nature*, vol. 456, no. 7218, pp. 53–59, 2008.
- [33] O. Morozova and M. A. Marra, "Applications of next-generation sequencing technologies in functional genomics," *Genomics*, vol. 92, no. 5, pp. 255–264, 2008.
- [34] T. J. Treangen and S. L. Salzberg, "Repetitive dna and next-generation sequencing: Computational challenges and solutions," *Nature Reviews Genetics*, vol. 13, no. 1, pp. 36–46, 2012.

- [35] N. Olova, F. Krueger, S. Andrews, *et al.*, “Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting dna methylation data,” *Genome biology*, vol. 19, no. 1, p. 33, 2018.
- [36] J. Eid, A. Fehr, J. Gray, *et al.*, “Real-time dna sequencing from single polymerase molecules,” *Science*, vol. 323, no. 5910, pp. 133–138, 2009.
- [37] A. Rhoads and K. F. Au, “Pacbio sequencing and its applications,” *Genomics, proteomics & bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015.
- [38] A. M. Wenger, P. Peluso, W. J. Rowell, *et al.*, “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome,” *Nature biotechnology*, vol. 37, no. 10, pp. 1155–1162, 2019.
- [39] D. Branton, D. W. Deamer, A. Marziali, *et al.*, “The potential and challenges of nanopore sequencing,” *Nature biotechnology*, vol. 26, no. 10, pp. 1146–1153, 2008.
- [40] D. Deamer, M. Akeson, and D. Branton, “Three decades of nanopore sequencing,” *Nature biotechnology*, vol. 34, no. 5, pp. 518–524, 2016.
- [41] T. Mantere, S. Kersten, and A. Hoischen, “Long-read sequencing emerging in medical genetics,” *Frontiers in genetics*, vol. 10, p. 432 668, 2019.
- [42] P. Biosciences. “Primrose: 5mc detection in pacbio hifi reads.” (2024), [Online]. Available: https://www.pacb.com/wp-content/uploads/poster_saunders.pdf (visited on 10/17/2025).
- [43] P. Ni, F. Nie, Z. Zhong, *et al.*, “Dna 5-methylcytosine detection and methylation phasing using pacbio circular consensus sequencing,” *Nature communications*, vol. 14, no. 1, p. 4054, 2023.
- [44] Y. Saletore, K. Meyer, J. Korlach, I. D. Vilfan, S. Jaffrey, and C. E. Mason, “The birth of the epitranscriptome: Deciphering the function of rna modifications,” *Genome biology*, vol. 13, no. 10, p. 175, 2012.
- [45] A. Magi, G. Mattei, A. Mingrino, *et al.*, “Gasoline: Detecting germline and somatic structural variants from long-reads data,” *Scientific Reports*, vol. 13, no. 1, p. 20 817, 2023.

- [46] S. A. McCarroll, A. Huett, P. Kuballa, *et al.*, “Deletion polymorphism upstream of *irgm* associated with altered *irgm* expression and crohn’s disease,” *Nature genetics*, vol. 40, no. 9, pp. 1107–1112, 2008.
- [47] T.-L. Yang, X.-D. Chen, Y. Guo, *et al.*, “Genome-wide copy-number-variation study identified a susceptibility gene, *ugt2b17*, for osteoporosis,” *The American Journal of Human Genetics*, vol. 83, no. 6, pp. 663–674, 2008.
- [48] M. R. Cosenza, B. Rodriguez-Martin, and J. O. Korbel, “Structural variation in cancer: Role, prevalence, and mechanisms,” *Annual Review of Genomics and Human Genetics*, vol. 23, no. 1, pp. 123–152, 2022.
- [49] Y. Li, N. D. Roberts, J. A. Wala, *et al.*, “Patterns of somatic structural variation in human cancer genomes,” *Nature*, vol. 578, no. 7793, pp. 112–121, 2020.
- [50] G. Escaramís, E. Docampo, and R. Rabionet, “A decade of structural variants: Description, history and methods to detect structural variation,” *Briefings in functional genomics*, vol. 14, no. 5, pp. 305–314, 2015.
- [51] J. M. Levisky and R. H. Singer, “Fluorescence in situ hybridization: Past, present and future,” *Journal of cell science*, vol. 116, no. 14, pp. 2833–2838, 2003.
- [52] L. E. Vissers, B. B. de Vries, K. Osoegawa, *et al.*, “Array-based comparative genomic hybridization for the genomewide detection of sub-microscopic chromosomal abnormalities,” *The American Journal of Human Genetics*, vol. 73, no. 6, pp. 1261–1270, 2003.
- [53] L. Tattini, R. D’Aurizio, and A. Magi, “Detection of genomic structural variants from next-generation sequencing data,” *Frontiers in bioengineering and biotechnology*, vol. 3, p. 92, 2015.
- [54] J. Huddleston, M. J. Chaisson, K. M. Steinberg, *et al.*, “Discovery and genotyping of structural variation from long-read haploid genome sequence data,” *Genome research*, vol. 27, no. 5, pp. 677–685, 2017.
- [55] S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani, “Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing,” *Genome biology*, vol. 20, no. 1, p. 117, 2019.

- [56] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck, “Structural variant calling: The long and the short of it,” *Genome biology*, vol. 20, no. 1, p. 246, 2019.
- [57] M. U. Ahsan, Q. Liu, J. E. Perdomo, L. Fang, and K. Wang, “A survey of algorithms for the detection of genomic structural variants from long-read sequencing data,” *Nature Methods*, vol. 20, no. 8, pp. 1143–1158, 2023.
- [58] M. Smolka, L. F. Paulin, C. M. Grochowski, *et al.*, “Detection of mosaic and population-level structural variants with sniffles2,” *Nature biotechnology*, vol. 42, no. 10, pp. 1571–1580, 2024.
- [59] T. Jiang, Y. Liu, Y. Jiang, *et al.*, “Long-read-based human genomic structural variation detection with cutesv,” *Genome biology*, vol. 21, no. 1, p. 189, 2020.
- [60] D. Heller and M. Vingron, “Svim: Structural variant identification using mapped long reads,” *Bioinformatics*, vol. 35, no. 17, pp. 2907–2915, 2019.
- [61] D. Bolognini and A. Magi, “Evaluation of germline structural variant calling methods for nanopore sequencing data,” *Frontiers in Genetics*, vol. 12, p. 761 791, 2021.
- [62] J. Lin, P. Jia, S. Wang, W. Kusters, and K. Ye, “Comparison and benchmark of structural variants detected from long read and long-read assembly,” *Briefings in Bioinformatics*, vol. 24, no. 4, bbad188, 2023.
- [63] N. Dierckxsens, T. Li, J. R. Vermeesch, and Z. Xie, “A benchmark of structural variation detection by long reads through a realistic simulated model,” *Genome biology*, vol. 22, no. 1, p. 342, 2021.
- [64] D. Eppstein, M. Löffler, and D. Strash, “Listing all maximal cliques in sparse graphs in near-optimal time,” in *International symposium on algorithms and computation*, Springer, 2010, pp. 403–414.
- [65] C. Chiang, R. M. Layer, G. G. Faust, *et al.*, “Speedseq: Ultra-fast personal genome analysis and interpretation,” *Nature methods*, vol. 12, no. 10, pp. 966–968, 2015.
- [66] Y. Ono, K. Asai, and M. Hamada, “Pbsim2: A simulator for long-read sequencers with a novel generative model of quality scores,” *Bioinformatics*, vol. 37, no. 5, pp. 589–595, 2021.

- [67] H. Li, “Minimap2: Pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.
- [68] F. J. Sedlazeck, P. Rescheneder, M. Smolka, *et al.*, “Accurate detection of complex structural variations using single-molecule sequencing,” *Nature methods*, vol. 15, no. 6, pp. 461–468, 2018.
- [69] J. M. Zook, N. F. Hansen, N. D. Olson, *et al.*, “A robust benchmark for detection of germline large deletions and insertions,” *Nature biotechnology*, vol. 38, no. 11, pp. 1347–1355, 2020.
- [70] D. C. Jeffares, C. Jolly, M. Hoti, *et al.*, “Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast,” *Nature communications*, vol. 8, no. 1, p. 14061, 2017.
- [71] J. E. Valle-Inclan, N. J. Besselink, E. de Bruijn, *et al.*, “A multi-platform reference for somatic structural variation detection,” *Cell genomics*, vol. 2, no. 6, 2022.
- [72] J. R. Belyeu, M. Chowdhury, J. Brown, *et al.*, “Samplot: A platform for structural variant visual validation and automated filtering,” *Genome biology*, vol. 22, no. 1, p. 161, 2021.
- [73] G. Mattei, M. Baragli, B. Gega, *et al.*, “Poremeth2 for decoding the evolution of methylome alterations with nanopore sequencing,” *Genome Research*, vol. 35, no. 11, pp. 2501–2512, 2025.
- [74] H. I. Roach, N. Yamada, K. S. Cheung, *et al.*, “Association between the abnormal expression of matrix-degrading enzymes by human osteoarthritic chondrocytes and demethylation of specific cpG sites in the promoter regions,” *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 52, no. 10, pp. 3110–3124, 2005.
- [75] K. Hashimoto, S. Kokubun, E. Itoi, and H. I. Roach, “Improved quantification of dna methylation using methylation-sensitive restriction enzymes and real-time pcr,” *Epigenetics*, vol. 2, no. 2, pp. 86–91, 2007.
- [76] M. Stevens, J. B. Cheng, D. Li, *et al.*, “Estimating absolute methylation levels at single-cpg resolution from methylation enrichment and restriction enzyme sequencing methods,” *Genome research*, vol. 23, no. 9, pp. 1541–1553, 2013.

- [77] F. V. Jacinto, E. Ballestar, and M. Esteller, "Methyl-dna immunoprecipitation (medip): Hunting down the dna methylome," *Biotechniques*, vol. 44, no. 1, pp. 35–43, 2008.
- [78] D. Serre, B. H. Lee, and A. H. Ting, "Mbd-isolated genome sequencing provides a high-throughput and comprehensive survey of dna methylation in the human genome," *Nucleic acids research*, vol. 38, no. 2, pp. 391–399, 2010.
- [79] M. Frommer, L. E. McDonald, D. S. Millar, *et al.*, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 5, pp. 1827–1831, 1992.
- [80] R. A. Harris, T. Wang, C. Coarfa, *et al.*, "Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications," *Nature biotechnology*, vol. 28, no. 10, pp. 1097–1105, 2010.
- [81] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, "Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis," *Nucleic acids research*, vol. 33, no. 18, pp. 5868–5877, 2005.
- [82] K. Tanaka and A. Okamoto, "Degradation of dna by bisulfite treatment," *Bioorganic & medicinal chemistry letters*, vol. 17, no. 7, pp. 1912–1915, 2007.
- [83] M. J. Ziller, K. D. Hansen, A. Meissner, and M. J. Aryee, "Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing," *Nature methods*, vol. 12, no. 3, pp. 230–232, 2015.
- [84] Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, and A. Rao, "The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing," *PloS one*, vol. 5, no. 1, e8888, 2010.
- [85] M. Bibikova, B. Barnes, C. Tsan, *et al.*, "High density dna methylation array with single cpg site resolution," *Genomics*, vol. 98, no. 4, pp. 288–295, 2011.
- [86] M. Bibikova, J. Le, B. Barnes, *et al.*, "Genome-wide dna methylation profiling using infinium® assay," *Epigenomics*, vol. 1, no. 1, pp. 177–200, 2009.

- [87] S. Moran, C. Arribas, and M. Esteller, "Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences," *Epigenomics*, vol. 8, no. 3, pp. 389–399, 2016.
- [88] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, "Opportunities and challenges in long-read sequencing data analysis," *Genome biology*, vol. 21, no. 1, p. 30, 2020.
- [89] Q. Gouil and A. Keniry, "Latest techniques to study dna methylation," *Essays in biochemistry*, vol. 63, no. 6, pp. 639–648, 2019.
- [90] W. A. Cheung, A. F. Johnson, W. J. Rowell, *et al.*, "Direct haplotype-resolved 5-base hifi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort," *Nature Communications*, vol. 14, no. 1, p. 3090, 2023.
- [91] J. T. Simpson, R. E. Workman, P. Zuzarte, M. David, L. Dursi, and W. Timp, "Detecting dna cytosine methylation using nanopore sequencing," *Nature methods*, vol. 14, no. 4, pp. 407–410, 2017.
- [92] M. Stoiber, J. Quick, R. Egan, *et al.*, "De novo identification of dna modifications enabled by genome-guided nanopore signal processing," *BioRxiv*, p. 094672, 2016.
- [93] P. Ni, N. Huang, Z. Zhang, *et al.*, "Deepsignal: Detecting dna methylation state from nanopore sequencing reads using deep-learning," *Bioinformatics*, vol. 35, no. 22, pp. 4586–4595, 2019.
- [94] V. Akbari, J.-M. Garant, K. O'Neill, *et al.*, "Megabase-scale methylation phasing using nanopore long reads and nanomethphase," *Genome biology*, vol. 22, no. 1, p. 68, 2021.
- [95] Y. Park and H. Wu, "Differential methylation analysis for bs-seq data under general experimental design," *Bioinformatics*, vol. 32, no. 10, pp. 1446–1453, 2016.
- [96] R. Snajder, A. Leger, O. Stegle, and M. J. Bonder, "Pycometh: A toolbox for differential methylation testing from nanopore methylation calls," *Genome biology*, vol. 24, no. 1, p. 83, 2023.
- [97] A. Magi, G. Mattei, A. Mingrino, *et al.*, "High-resolution nanopore methylome-maps reveal random hyper-methylation at cpg-poor regions as driver of chemoresistance in leukemias," *Communications Biology*, vol. 6, no. 1, p. 382, 2023.

- [98] A. Magi, M. Benelli, G. Marseglia, G. Nannetti, M. R. Scordo, and F. Torricelli, "A shifting level model algorithm that identifies aberrations in array-cgh data," *Biostatistics*, vol. 11, no. 2, pp. 265–280, 2010.
- [99] H. Xie, M. Wang, A. De Andrade, *et al.*, "Genome-wide quantitative assessment of variation in dna methylation patterns," *Nucleic acids research*, vol. 39, no. 10, pp. 4099–4108, 2011.
- [100] J. M. Bhasin and A. H. Ting, "Goldmine integrates information placing genomic ranges into meaningful biological contexts," *Nucleic acids research*, vol. 44, no. 12, pp. 5550–5556, 2016.
- [101] S. Heinz, C. Benner, N. Spann, *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities," *Molecular cell*, vol. 38, no. 4, pp. 576–589, 2010.
- [102] K. Kupkova, J. V. Mosquera, J. P. Smith, *et al.*, "Genomicdistributions: Fast analysis of genomic intervals with bioconductor," *BMC genomics*, vol. 23, no. 1, p. 299, 2022.
- [103] R. G. Cavalcante and M. A. Sartor, "Annotatr: Genomic regions in context," *Bioinformatics*, vol. 33, no. 15, pp. 2381–2383, 2017.
- [104] J. G. Tate, S. Bamford, H. C. Jubb, *et al.*, "Cosmic: The catalogue of somatic mutations in cancer," *Nucleic acids research*, vol. 47, no. D1, pp. D941–D947, 2019.
- [105] S. A. Lambert, A. Jolma, L. F. Campitelli, *et al.*, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.
- [106] Y.-Y. Wang, W.-H. Chen, P.-P. Xiao, *et al.*, "Gear: A database of genomic elements associated with drug resistance," *Scientific reports*, vol. 7, no. 1, p. 44085, 2017.
- [107] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.
- [108] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature methods*, vol. 14, no. 4, pp. 417–419, 2017.
- [109] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

- [110] Y. Liao, G. K. Smyth, and W. Shi, "The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads," *Nucleic acids research*, vol. 47, no. 8, e47–e47, 2019.
- [111] B. L. Bass, "Rna editing by adenosine deaminases that act on rna," *Annual review of biochemistry*, vol. 71, no. 1, pp. 817–846, 2002.
- [112] K. Nishikura, "Functions and regulation of rna editing by adar deaminases," *Annual review of biochemistry*, vol. 79, no. 1, pp. 321–349, 2010.
- [113] R. Pecori, S. Di Giorgio, J. Paulo Lorenzo, and F. Nina Papavasiliou, "Functions and consequences of aid/apobec-mediated dna and rna deamination," *Nature Reviews Genetics*, vol. 23, no. 8, pp. 505–518, 2022.
- [114] S. M. Rueter, T. R. Dawson, and R. B. Emeson, "Regulation of alternative splicing by rna editing," *Nature*, vol. 399, no. 6731, pp. 75–80, 1999.
- [115] B. Sommer, M. Köhler, R. Sprengel, and P. H. Seeburg, "Rna editing in brain controls a determinant of ion flow in glutamate-gated channels," *Cell*, vol. 67, no. 1, pp. 11–19, 1991.
- [116] N. M. Mannion, S. M. Greenwood, R. Young, *et al.*, "The rna-editing enzyme adar1 controls innate immune responses to rna," *Cell reports*, vol. 9, no. 4, pp. 1482–1494, 2014.
- [117] F. Galeano, S. Tomaselli, F. Locatelli, and A. Gallo, "A-to-i rna editing: The "adar" side of human cancer," in *Seminars in cell & developmental biology*, Elsevier, vol. 23, 2012, pp. 244–250.
- [118] C. R. Walkley and J. B. Li, "Rewriting the transcriptome: Adenosine-to-inosine rna editing by adars," *Genome biology*, vol. 18, no. 1, p. 205, 2017.
- [119] A. Cappannini, A. Ray, E. Purta, *et al.*, "Modomics: A database of rna modifications and related information. 2023 update," *Nucleic Acids Research*, vol. 52, no. D1, pp. D239–D244, 2024.
- [120] X. Wang, Z. Lu, A. Gomez, *et al.*, "N 6-methyladenosine-dependent regulation of messenger rna stability," *Nature*, vol. 505, no. 7481, pp. 117–120, 2014.

- [121] H. Shi, P. Chai, R. Jia, and X. Fan, "Novel insight into the regulatory roles of diverse rna modifications: Re-defining the bridge between transcription and translation," *Molecular cancer*, vol. 19, no. 1, p. 78, 2020.
- [122] D. Wiener and S. Schwartz, "The epitranscriptome beyond m6a," *Nature Reviews Genetics*, vol. 22, no. 2, pp. 119–131, 2021.
- [123] V. Griesche, S. Di Giorgio, J. Brettschneider, *et al.*, "M6a regulates adar1-mediated rna editing during macrophage activation," *bioRxiv*, pp. 2025–01, 2025.
- [124] J.-F. Xiang, Q. Yang, C.-X. Liu, M. Wu, L.-L. Chen, and L. Yang, "N6-methyladenosines modulate a-to-i rna editing," *Molecular cell*, vol. 69, no. 1, pp. 126–135, 2018.
- [125] R. Desrosiers, K. Friderici, and F. Rottman, "Identification of methylated nucleosides in messenger rna from novikoff hepatoma cells," *Proceedings of the National Academy of Sciences*, vol. 71, no. 10, pp. 3971–3975, 1974.
- [126] N. Liu, M. Parisien, Q. Dai, G. Zheng, C. He, and T. Pan, "Probing n6-methyladenosine rna modification status at single nucleotide resolution in mrna and long noncoding rna," *Rna*, vol. 19, no. 12, pp. 1848–1856, 2013.
- [127] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, *et al.*, "Topology of the human and mouse m6a rna methylomes revealed by m6a-seq," *Nature*, vol. 485, no. 7397, pp. 201–206, 2012.
- [128] M. Schaefer, T. Pollex, K. Hanna, and F. Lyko, "Rna cytosine methylation analysis by bisulfite sequencing," *Nucleic acids research*, vol. 37, no. 2, e12–e12, 2009.
- [129] J. E. Squires, H. R. Patel, M. Nousch, *et al.*, "Widespread occurrence of 5-methylcytosine in human coding and non-coding rna," *Nucleic acids research*, vol. 40, no. 11, pp. 5023–5033, 2012.
- [130] T. M. Carlile, M. F. Rojas-Duran, and W. V. Gilbert, "Pseudo-seq: Genome-wide detection of pseudouridine modifications in rna," in *Methods in enzymology*, vol. 560, Elsevier, 2015, pp. 219–245.
- [131] S. Schwartz, D. A. Bernstein, M. R. Mumbach, *et al.*, "Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncna and mrna," *Cell*, vol. 159, no. 1, pp. 148–162, 2014.

- [132] M. Sakurai, T. Yano, H. Kawabata, H. Ueda, and T. Suzuki, "Inosine cyanoethylation identifies a-to-i rna editing sites in the human transcriptome," *Nature chemical biology*, vol. 6, no. 10, pp. 733–740, 2010.
- [133] D. R. Garalde, E. A. Snell, D. Jachimowicz, *et al.*, "Highly parallel direct rna sequencing on an array of nanopores," *Nature methods*, vol. 15, no. 3, pp. 201–206, 2018.
- [134] R. E. Workman, A. D. Tang, P. S. Tang, *et al.*, "Nanopore native rna sequencing of a human poly (a) transcriptome," *Nature methods*, vol. 16, no. 12, pp. 1297–1305, 2019.
- [135] M. T. Parker, K. Knop, A. V. Sherwood, *et al.*, "Nanopore direct rna sequencing maps the complexity of arabidopsis mrna processing and m6a modification," *elife*, vol. 9, e49658, 2020.
- [136] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, "Nanopore sequencing technology, bioinformatics and applications," *Nature biotechnology*, vol. 39, no. 11, pp. 1348–1365, 2021.
- [137] P. N. Pratanwanich, F. Yao, Y. Chen, *et al.*, "Identification of differential rna modifications from nanopore direct rna sequencing with xpore," *Nature biotechnology*, vol. 39, no. 11, pp. 1394–1402, 2021.
- [138] C. Soneson, Y. Yao, A. Bratus-Neuenschwander, A. Patrignani, M. D. Robinson, and S. Hussain, "A comprehensive examination of nanopore native rna sequencing for characterization of complex transcriptomes," *Nature communications*, vol. 10, no. 1, p. 3359, 2019.
- [139] M. Jain, R. Abu-Shumays, H. E. Olsen, and M. Akeson, "Advances in nanopore direct rna sequencing," *Nature methods*, vol. 19, no. 10, pp. 1160–1164, 2022.
- [140] W. Stephenson, R. Razaghi, S. Busan, K. M. Weeks, W. Timp, and P. Smibert, "Direct detection of rna modifications and structure using single-molecule nanopore sequencing," *Cell genomics*, vol. 2, no. 2, 2022.
- [141] A. Leger, P. P. Amaral, L. Pandolfini, *et al.*, "Rna modifications detection by comparative nanopore direct rna sequencing," *Nature communications*, vol. 12, no. 1, p. 7198, 2021.

- [142] P. Jenjaroenpun, T. Wongsurawat, T. D. Wadley, *et al.*, “Decoding the epitranscriptional landscape from native rna sequences,” *Nucleic acids research*, vol. 49, no. 2, e7–e7, 2021.
- [143] Y. Zou, M. U. Ahsan, J. Chan, *et al.*, “A comparative evaluation of computational models for rna modification detection using nanopore sequencing with rna004 chemistry,” *bioRxiv*, 2025.
- [144] G. Diensthuber, I. Milenkovic, L. Llovera, A. Milovanovic, F. Pelizzari, and E. M. Novoa, “Systematic benchmarking of basecalling models for rna modification detection with highly-multiplexed nanopore sequencing,” *bioRxiv*, pp. 2025–07, 2025.
- [145] H. Liu, O. Begik, M. C. Lucas, *et al.*, “Accurate detection of m6a rna modifications in native rna sequences,” *Nature communications*, vol. 10, no. 1, p. 4079, 2019.
- [146] H. Liu, O. Begik, and E. M. Novoa, “Epinano: Detection of m6a rna modifications using oxford nanopore direct rna sequencing,” in *RNA modifications: methods and protocols*, Springer, 2021, pp. 31–52.
- [147] M. Lemercier, P. Arrubarrena, S. Di Giorgio, *et al.*, “Path signatures enable model-free mapping of rna modifications,” *arXiv preprint arXiv:2511.08855*, 2025.
- [148] L. Cozzuto, H. Liu, L. P. Pryszcz, *et al.*, “Masterofpores: A workflow for the analysis of oxford nanopore direct rna sequencing datasets,” *Frontiers in genetics*, vol. 11, p. 211, 2020.
- [149] L. Cozzuto, A. Delgado-Tejedor, T. Hermoso Pulido, E. M. Novoa, and J. Ponomarenko, “Nanopore direct rna sequencing data processing and analysis using masterofpores,” in *Computational Epigenomics and Epitranscriptomics*, Springer, 2023, pp. 185–205.
- [150] O. N. Technologies. “Dorado documentation.” (2025), [Online]. Available: https://software-docs.nanoporetech.com/dorado/latest/basecaller/move_table/.
- [151] H. Gamaarachchi, C. W. Lam, G. Jayatilaka, *et al.*, “Gpu accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis,” *BMC bioinformatics*, vol. 21, no. 1, p. 343, 2020.
- [152] S. Kovaka, P. W. Hook, K. M. Jenike, *et al.*, “Uncalled4 improves nanopore dna and rna modification detection via fast and accurate signal alignment,” *Nature Methods*, pp. 1–11, 2025.