Genome Analysis Methods using Long Read Nanopore Sequencing

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

Fachbereich Mathematik und Informatik

Institut für Informatik

Freie Universität Berlin

vorgelegt von

Pay Giesselmann

Berlin 2021

Erstgutachter:	Prof. Dr. Knut Reinert Freie Universiät Berlin Berlin
Zweitgutachter:	Prof. Dr. Peter F. Stadler Universität Leipzig Leipzig
Disputation:	12.10.2021

M.Eng. Pay Giesselmann
Genome Analysis Methods using Long Read Nanopore Sequencing
Dissertation, 01.12.2021

Max-Planck-Institue for Molecular Genetics
Department of Genome Regulation
Meissner Lab
Ihnestraße 63-73
14195 Berlin

Freie Universiät Berlin
Fachbereich Mathematik und Informatik
Institut für Informatik
Reinert Lab
Takustraße 9
14195 Berlin

Abstract

Third-generation long-read technologies denote the latest progression in high throughput DNA and RNA sequence analysis. Complementing the widespread second-generation short-read platforms, long-read sequencing adds unique application opportunities by generating previously unattainable read lengths. Despite the remaining higher error rate compared to short reads, single-molecule real-time sequencing (SMRT) and nanopore sequencing advanced to be state-of-the-art for *de-novo* genome assemblies and identification of structural variants. Continuous throughput and accuracy improvements lead to development of novel methods and applications at a fast pace. We identify major application fields and key bioinformatic software for long-read sequencing data analysis by employing a data driven literature research. The integration of citations and keywords into a literature graph provides a scaling approach to analyze an exponentially growing number of third-generation sequencing related publications. Even though sparking the development of countless bioinformatics software, the streamlined nanopore data processing into standardized formats is still lacking. As an enabling step for its successful application, we developed *Nanopype*, a modular and scalable pipeline. Our approach facilitates the basic steps of basecalling, alignment, methylation- and structural variant detection with exchangeable tools in each module. Optimized for the usage on high performance compute clusters, we propose a raw data management, capable of handling multiple sequencing devices placed locally and remotely. Strict version control of integrated tools and deployment as containerized software, ensure reproducibility across projects and laboratories. Finally, we analyze disease associated repeat regions utilizing targeted nanopore sequencing and the *Nanopype* processing infrastructure. The expansion of unstable genomic short tandem repeats (STRs) is of particular interest as it causes more than 30 Mendelian human disorders. Long stretches of repetitive sequence render these regions inaccessible for short-read sequencing by synthesis. Furthermore, finding current nanopore basecalling algorithms insufficient to resolve the repeat length, we developed *STRique*, a raw nanopore signal based repeat detection and quantification software. We demonstrate the precise analysis of repeat lengths on patient-derived samples containing C9orf72 and FMR1 repeat expansions. The additional integration of repeat- and nearby promoter-methylation levels reveal a repeat length depending gain, suggesting an epigenetic response to the expansion. Taken together, this work contributes to further increase the usability and provides novel insights based on third-generation nanopore sequencing.

Zusammenfassung

Im Bereich der DNA und RNA-Sequenzierung stellen Nanopore Technologien den neusten Fortschritt da. Die Sequenzierung von deutlich längeren Fragmenten eröffnet einzigartige Anwendungsmöglichkeiten im Vergleich zu den weit verbreiteten, synthese-basierten Systemen von zum Beispiel *Illumina*. Kommerziell verfügbare Plattformen werden zur Zeit von *Pacific Biosciences* (PacBio) und *Oxford Nanopore Technologies* (ONT) vertrieben. Ungeachtet der höheren Fehlerrate im Vergleich zu bisherigen Systemen hat sich die Nanopore-Sequenzierung zum Stand der Technik für Genom-Assemblierung und zur Identifikation von Strukturvarianten entwickelt. Das direkte Auslesen chemischer Basen-Modifikationen, insbesondere von 5-Methylcytosin, ermöglicht die Untersuchung von bisher schwer zugänglichen Regionen eines Genoms oder die Verknüpfung von entfernten Merkmalen auf einzelnen Molekülen, was den Einsatz der Nanopore-Sequenzierung in der Epigenetik attraktiv macht. Eine kontinuierliche Verbesserungen des Durchsatzes und der Genauigkeit führen derzeit zu einer rasanten Entwicklung neuer Methoden und Anwendungen. Mit Hilfe einer Metadaten basierten Literaturrecherche werden zunächst wichtige Anwendungsfelder und Softwarelösungen für die Analyse von Nanopore Sequenzierdaten identifiziert. Die Integration von Zitationen und Schlüsselwörtern in einen Literaturgraph bietet einen skalierenden Ansatz, um die exponentiell wachsende Anzahl von Publikationen zu analysieren. Obwohl die Entwicklung unzähliger Analyseprogramme vorangetrieben wurde, mangelt es immer noch an einer effizienten Verarbeitung von Nanopore-Daten mit standardisierten Dateiformaten. Als Voraussetzung für eine erfolgreiche Anwendung haben wir daher zunächst *Nanopype* entwickelt, eine modulare und skalierbare Pipeline. Unser Ansatz ermöglicht es, die grundlegenden Schritte Basecalling, Alignment, Methylierungs- und Variationsdetektierung mit austauschbaren Tools in jedem Schritt durchzuführen. Optimiert für den Einsatz auf Hochleistungs-Rechenclustern, wird zudem ein Rohdatenmanagement vorgeschlagen, das in der Lage ist, mehrere lokal und entfernt platzierte Sequenziergeräte zu verwalten. Eine strikte Versionskontrolle der integrierten Tools und die Bereitstellung als Softwarecontainer gewährleisten die Reproduzierbarkeit über Projekte und Labore hinweg. Schließlich analysieren wir krankheitsassoziierte Variationen genomische Regionen unter Verwendung der Nanopore-Technologie und der Infrastruktur von *Nanopype*. Die Ausweitung von instabilen kurzen Tandemwiederholungen (Short-Tandem-Repeats, STRs) ist von besonderem Interesse, da sie mehr als 30 menschliche Erkrankungen verursacht. Lange Abschnitte der repetitiven

Sequenz machen diese Regionen unzugänglich für kurze Fragmente aus einer Sequenzierung durch Synthese. Da die derzeitigen Nanopore-Basecalling-Algorithmen ebenfalls unzureichend sind, um die exakte Wiederholungsanzahl aufzulösen, haben wir *STRique* entwickelt, eine auf dem Nanopore-Rohsignal aufbauende Software zur Erkennung und Quantifizierung von Wiederholungen. Wir demonstrieren die präzise Bestimmung von Wiederholungslängen an Patientenproben, die C9orf72- und FMR1-Expansionen enthalten. Ein Zusammenhang zwischen Wiederholungszahl und der erhöhten Methylierung des nahegelegenen C9orf72 Promoter deutet auf eine epigenetische Reaktion auf die Expansion hin. Zusammengefasst trägt diese Arbeit dazu bei, die generelle Anwendbarkeit der Nanopore-Sequenzierung weiter zu verbessern und demonstriert eine Analyse von repetitiven genomischen Regionen auf Basis des Rohsignals, die in dieser Form mit bisherigen Methoden nicht möglich ist.

Bibliographic Information

This thesis is based on the following publications:

Pay Giesselmann, Sara Hetzel, Franz-Josef Müller, Alexander Meissner and Helene Kretzmer. *Nanopype: a modular and scalable nanopore data processing pipeline*. **Bioinformatics**, Volume 35, Issue 22, 15 November 2019, Pages 4770–4772.
https://doi.org/10.1093/bioinformatics/btz461

Pay Giesselmann*, Björn Brändl*, Etienne Raimondeau, Rebecca Bowen, Christian Rohrandt, Rashmi Tandon, Helene Kretzmer, Günter Assum, Christina Galonska, Reiner Siebert, Ole Ammerpohl, Andrew Heron, Susanne A. Schneider, Julia Ladewig, Philipp Koch, Bernhard M. Schuldt, James E. Graham, Alexander Meissner and Franz-Josef Müller. *Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing*. **Nature Biotechnology**, Volume 37, 18 November 2019, Pages 1478–1481.
https://doi.org/10.1038/s41587-019-0293-x

*joint first authorship

Acknowledgement

Contents

1 Introduction

It is astonishing how cells in an organism can differentiate into a variety of specialized types with a multitude of different tasks despite being based on the same genetic information coded into the DNA. Beginning with a single cell, a temporal, spatial and functional coordination determines the growth and body formation of eukaryotic organisms. Depending on the cell state, only a fraction of genes is actively transcribed in a given cell while others remain repressed. Acting as an additional regulatory layer, the epigenome describes a set of chemical modifications made to the DNA, controlling, among other things, the activation and transcription of genes into RNA, and ultimately, proteins. Current sequencing technologies and their connected bioinformatics provide researchers with the tools to study DNA sequence and epigenetic state down to single-cell and single-molecule resolution. Starting in 1990, and taking over a decade until completion, the human genome project incorporated an international team of researchers to decipher the first genetic blueprint to build a human being. Based on first-generation sequencing technology, the outcome was a nearly complete reference sequence, including gene annotations. Since then, the development of high-throughput, next-generation sequencing (NGS) technologies enabled studies of countless organisms, cell types and disease conditions. While being very reliable in terms of throughput and accuracy, sequenced fragments of a few hundreds nucleotides in length still limit the readout from repetitive regions or the resolution of long distance dependencies on a single-molecule.

1.1 Motivation

In the past five years, a third-generation of sequencing techniques is introducing new perspectives to the field of genome analysis by generating previously unattainable read lengths with averages in the tens of thousands of nucleotides. Under active development with frequent improvements, long-read sequencing provides new opportunities by, visually speaking, increasing the size of the puzzle pieces. Moreover, direct sequencing of DNA and RNA molecules using the nanopore technology facilitates the detection of different base modifications. Nanopore sequencing can be used to produce large, multifaceted data sets within a few days, and therefore provides

the opportunity for a new field of long read sequencing research at the intersection of genomics, computer science and engineering. New data types, formats and error characteristics demand here the adaptation of existing, or even the development of new, algorithms for bioinformatic software.

1.2 Genome Regulation

Most cell types of an eukaryotic organism contain a copy of the genetic code in form of folded DNA in the nucleus. Virtually all mammals have diploid cells with homologous maternal and paternal copies, organized into chromosome pairs with the same genes at the same genomic locations. While the entire human genome is comprised of about three billion nucleotides in total, the longest continuous stretch of DNA is the first of 23 chromosome pairs with 247 million base pairs. To maintain its integrity and to support the chromatin structure formation, chromosomal DNA is wrapped around octamers of histone proteins (H2A, H2B, H3 and H4) called nucleosomes (Fig. 1.1).

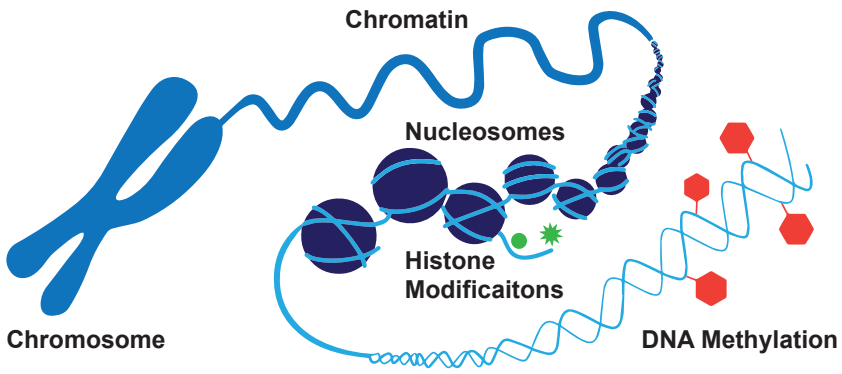


Fig. 1.1.: The DNA of each chromosome is wrapped around histones, forming nucleosomes and the chromatin structure. Chemical modifications to histone tails and individual base pairs impact properties and cellular function (Adapted from [1]).

The ~147bp of DNA directly wrapped around each nucleosome is protected against physical access from, for instance, transcription factors (TFs). The dynamic positioning of nucleosomes is therefore part of the epigenetic regulation process during transcription and replication. Accordingly, nucleosomes are enriched at compacted heterochromatin in mostly inactive genomic regions, while being thinned out at regulatory elements such as enhancers, insulators and transcribed genes [2]. A growing number of chemical modifications to histone tails is reported to impact inter-nucleosomal interactions, and also leads to recruitment of proteins and complexes involved in transcription, replication and DNA repair [3]. A well-characterized example is the methylation of lysine 79 on the H3 tail (H3K79me3) found at the tran-

scription start sites (TSS) of active genes [4]. In addition, the modification H3K4me3 marks active euchromatin, while H3K9me3 marks repressive heterochromatin. Gene expression is influenced by nearby active enhancer sites that are typically enriched with H3K27ac (acetylation). Finally, repressed and bivalent promoters are modified with H3K27me3 and the combination of H3K4me4 and H3K27me3, respectively.

For embryonic development in mammals, the methylation of cytosine in the CpG-context (5-methylcytosine, 5mC) is a vital epigenetic modification. Prominent examples of its importance include the X-chromosome inactivation in females and genomic imprinting. Furthermore, DNA methylation is associated with the silencing of transposable elements, shows high levels over actively transcribed gene bodies and correlates with gene repression at promoters [5]. In general, *de-novo* methylation of the DNA is primarily established by the DNMT3A and DNMT3B enzymes. During cell division, the methylation state on the nascent strand in the symmetric CpG-context is restored by the maintenance methyltransferase DNMT1. In the absence of DNMT1, the 5mC modification is therefore passively lost during each round of replication. However, can also be actively removed by oxidation through the TET methylcytosine dioxygenases. There is a side-effect of cytosine methylation in the form of potential spontaneous deamination of 5mC, resulting in inherited C to T transitions and ultimately, a depletion of CpG sites in mammalian genomes [6]. The human haploid genome for example, contains only around 29M CpGs instead of 188M as anticipated from the genome size. Thus, the conservation under evolutionary pressure underlines the importance of DNA methylation for mammals, while its diverse functions are still not fully understood.

Human somatic cells typically have a genome-wide mean methylation of approximately 70-80% [7]. In bulk methylation sequencing, millions of cells are measured at once, resulting in an averaging of the binary 5mC state of individual cells and yielding mean methylation levels per genomic position. The underlying distribution of methylation levels is not random, but follows a bimodal distribution, with CpG dense regions, also referred to as CpG islands (CGIs), being mostly unmethylated, while the genomic background of more isolated CpGs is mostly methylated [7]. Many CGIs in mammals act as promoters, however the majority remain unmethylated during differentiation even as gene silencing is driven by H3K27 methylation [5, 8]. In contrast, a well-studied example where DNA methylation is sufficient to repress transcription are imprinted genes, where either the maternal or paternal allele is transcribed, while the other allele remains silenced and methylated.

In general, the DNA methylation levels remain stable across cell divisions but show distinct patterns among different cell types. Aberrant DNA-methylation levels have been observed in different disease contexts, in particular with a heavy deregulation in cancer. While sharing a trend towards global hypomethylation and CGI hypermethylation compared to normal tissues (Fig. 1.2), different tumor types can still be identified, based only on their methylation footprint [9].

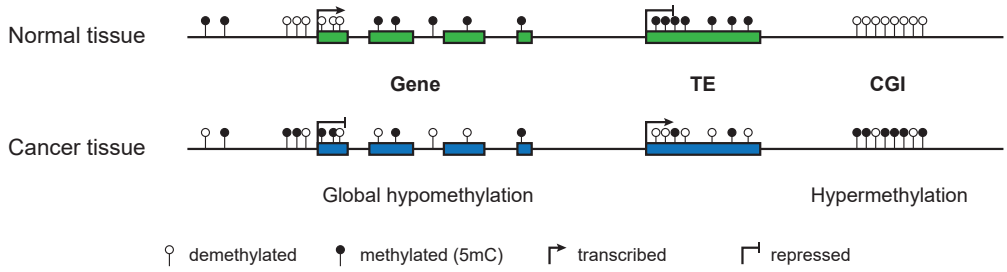


Fig. 1.2.: Deregulation of DNA methylation patterns in cancer tissues. A global loss across genes and transposable elements (TE) and local gain of methylation levels across CGIs is characteristic across cancer tissue types.

While several epigenetic regulators have been discovered and sequencing data for transcriptomes, histone modifications and methylation are available in abundance, the unknown function of high gene body methylation or similarities between extraembryonic tissue and altered methylation in cancer are only two examples of epigenetically regulated processes under active investigation [10].

1.3 Sequencing Technologies

"*For their contributions concerning the determination of base sequences in nucleic acids*", Walter Gilbert and Frederick Sanger received the 1980 Nobel Prize in Chemistry. Referred to as first-generation or Sanger sequencing, their 'dideoxy method' allowed the nucleotide sequence of an entire organism to be determined for the first time and with high accuracy [11]. While eventually superseded by second- and third-generation technologies, Sanger sequencing is still widely used as a method of verifying plasmid sequences or determining genotypes and sites of genome editing.

The development and commercialization of sequencing by synthesis approaches started in 2005 with the 454 Genome Sequencer and expanded from there to increase throughput and availability. These technologies are collectively referred to as next-generation sequencing (NGS) technologies. Different NGS platforms such as 454, Illumina, SOLiD and Ion torrent all rely on short reads of around a few hundred nucleotides in length. Illumina has long been an industry leader and was the first

company to reach the milestone of sequencing the human genome for only $1,000 USD with their HiSeq X Ten platform. Today, Illumina serves as the gold standard for second-generation sequencing [12].

1.3.1 Next Generation Sequencing

Illumina dye sequencing is a high throughput technology, reading millions of short DNA fragments in parallel and with high accuracy (>99.9%). As illustrated in Fig. 1.3, the process can be divided into library preparation, cluster amplification and the sequencing itself.

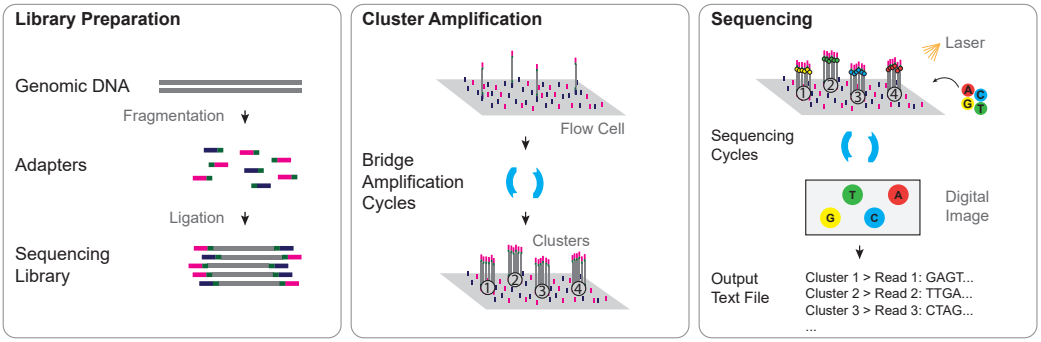


Fig. 1.3.: Next generation sequencing: The input DNA is fragmented and amplified into clusters of identical reads on a flow cell. Sequencing by synthesis determines the nucleotide sequence of all clusters simultaneously by measuring fluorescence of incorporated bases (Adapted from [13]).

In brief: Genomic input DNA is first sheared into fragments and ligated to distinct 5' and 3' end adapters. A polymerase chain reaction (PCR) amplification purifies the library for fragments with sequencing adapters on both ends. Subsequently, the double-stranded DNA is denatured and washed over a flow cell coated with short oligonucleotides, complementary to the sequencing adapters. The ends of single-stranded fragments ligate in sparse density with both, 5' and 3' adapter to the flow cell. During bridge amplification, the free adapters of single-stranded fragments are repeatedly bent to the flow cell surface and the complementary strand is synthesized. The denaturation of these double-stranded bridges results in a duplication of each template strand into its reverse complement. Lastly, one adapter type is cut away to unify the fragment orientation, resulting in dense clusters of identical single-stranded DNA fragments.

The actual sequencing process is divided into cycles, each detecting the respective next nucleotide of all clusters in parallel. During sequencing by synthesis, a polymerase assembles the complementary strand of each fragment from a mix of reversible terminating fluorescent nucleotides. A blocking group on every nucleotide

forces the polymerase to stop after each incorporation, allowing a camera to capture laser-excited fluorescence of all clusters. Different colors per nucleotide are recognized by the basecalling algorithm and concatenated into a sequence per cluster, referred to as a read. After unblocking and removal of the fluorophore, the process is repeated until the terminal read length is reached.

Limitations

Amplification during library preparation and the synthesis during sequencing are virtually error-free. However, the sequence quality, is affected by clusters running out of phase. Sporadic dropouts of the polymerase result in missing nucleotide incorporation within single fragments, leading to an increasing number of fragments lagging behind the cluster's synchronization. The resulting signal overlap is monitored by the basecaller and translated into a quality score per nucleotide and cluster. De-phsing of clusters currently limits the maximum sequencing length of NGS platforms to ~250nt.

With the exception of *de-novo* assemblies, a shared first processing step after genome wide sequencing is the alignment. After fragmentation and sequencing, the genomic origin of each read is initially unknown. An alignment algorithm determines all possible mappings of a read, commonly allowing for a certain degree of mismatches, insertions and deletions. Dissimilarities in read and reference sequence can result from either sequencing errors or differences between reference and the individual genome. At repetitive elements short reads may align to multiple genomic positions with the same edit distance. Depending on the application, filtering for unique alignments can therefore be necessary.

Sequencing the Epigenome High throughput second-generation sequencing has enabled a variety of protocols to be developed to extend their application beyond the readout of solely genomic sequence. The methods to detect histone modifications and DNA methylation by NGS are briefly outlined below.

Chromatin immunoprecipitation sequencing (ChIP-Seq) is a versatile genome-wide method to identify binding sites of DNA-associated proteins. Crosslinking of DNA-protein complexes fixes the current position of histones, for instance. First, a fragmentation step separates the DNA in protein-bound and unbound sections. Then, a modification specific magnetic antibody is used to immunoprecipitate the complex and pull down only the sequence fragments directly adjacent to a histone carrying

the epigenetic modification of interest. Sequencing and data analysis reveal histone modifications as peaks from local accumulations of reads.

Whole genome bisulfite sequencing (WGBS) is the state-of-the-art method to detect methylation at single base resolution on individual molecules [14]. A reaction with sodium bisulfite converts any unmethylated cytosine to uracil, while 5-methylcytosine remains unchanged and consequently encodes the methylation state into the sequence. Uracil and thymine are both amplified and sequenced as thymine (T), while only methylated cytosine is read as C. At genomic C positions, a read alignment containing a C indicates, therefore, a previously methylated site, while a mismatching T indicates an unmethylated site. The methylation state of individual reads is then summarized to a methylation rate per genomic position. A minimum coverage of 5X to 10X has been shown to be robust against inter-cell variability and allows the comparison across samples [15].

1.3.2 Third Generation Sequencing

In contrast to high-throughput short-read sequencing, third generation long-read technologies do not necessarily require amplification and yield read lengths in the range of tens of thousands of nucleotides. Single-molecule real-time sequencing (SMRT), introduced by Pacific Biosciences (PacBio) in 2011, is considered the first commercially available long-read technology [16]. While the idea of nanopore sequencing goes back to the 1980s, it took until 2014 to release the pocket sized MinION as the first device by Oxford Nanopore Technologies (ONT) [17].

SMRT is based on individual DNA fragments fixed into zero-mode waveguides (ZMW). For the library preparation, genomic input DNA is fragmented to typically 8-15kb and ligated to hairpin adapters. Termed SMRTbell, each molecule is denatured and as circular single stranded DNA fixed into the ZMWs on the flow cell. During the sequencing process, a polymerase located at the transparent bottom of each well synthesizes the complementary strand from nucleotides labeled with a fluorescent dye. At a speed of 10nt/s, light impulses from laser-excited fluorescence are the primary measurement to determine the nucleotide sequence of each read (Fig. 1.4, left box). According to the manufacturer, a Sequel IIe flow cell with 8M ZMWs can generate up to 4M reads during 30h of sequencing. [1] More recently, PacBio advanced to high-fidelity (HiFi) reads, which are generated as consensus sequence from multiple passes of the polymerase around the template strand. HiFi reads increase the single read accuracy to >99% at the cost of the overall read length. Without

[1] https://www.pacb.com/products-and-services/sequel-system, accessed 12/2020

PCR amplification of the input material, epigenetic base modifications remain on the sample DNA and impact the output signal as extended pauses between light impulses. The most prominent example is the detection of bacterial 6-methyladenine (6mA). However, the sensitivity to 5mC remains a proof of concept and studies reporting its successful application are lacking.

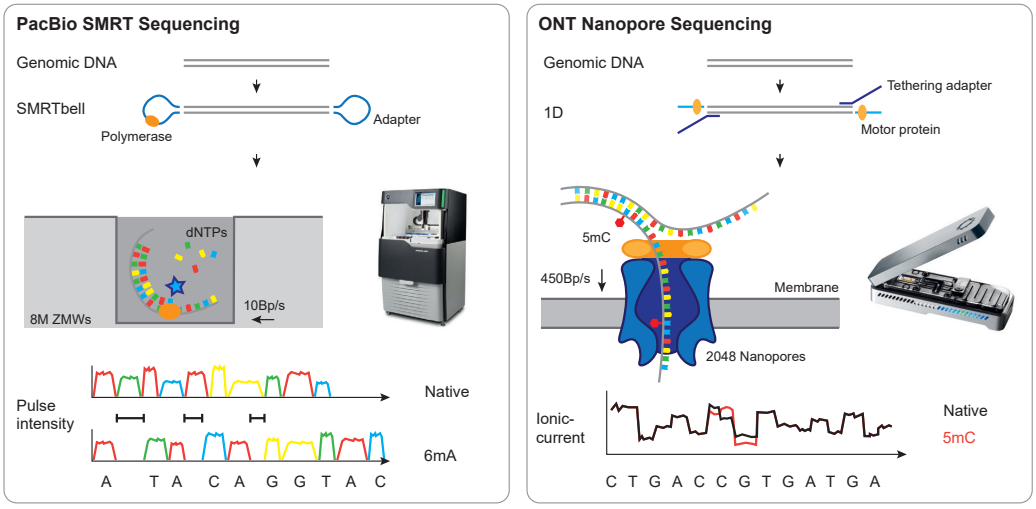


Fig. 1.4.: PacBio SMRT and ONT nanopore sequencing: Single-molecule real-time sequencing by circular synthesis of secondary strands measured as laser-excited fluorescence pulses. Nanopore sequencing of DNA and RNA strands measured as changes in ionic currents while traversing through the pore. Base modifications can be directly detected as either extended pulse pauses or characteristic current level differences.

Both NGS and SMRT sequencing technologies utilize the fluorescence signal of nucleotides synthesized into a complementary strand of the sequenced DNA fragment. Nanopore sequencing follows a fundamentally different principle. The sequencing is no longer based on the synthesis of a second strand, as individual pores asynchronously read multiple molecules one after the other, in contrast to the previous parallel approaches. Additionally, nanopore sequencing is not limited to processing only double stranded DNA, but also directly reads single stranded RNA. The nanopore flow cell is built of pores embedded into a membrane separating two chambers. After loading the flow cell with a running buffer, a voltage applied over the membrane causes a constant ionic current through each pore. With impact in the range of pico ampere, any molecule passing through the pore reduces the level of the ionic current depending on its physical and chemical properties.

Initial steps of the library preparation like fragmentation, end-repair and size selection are shared across long-read technologies. For nanopore sequencing, a motor protein and a tethering adapter are ligated to both ends of the input DNA. The tethering adapter helps to guide the reads to pores on the flow-cell, where the motor protein is attaching to the pore and controlling the sequencing speed to ~450nt/s for

DNA or ~70nt/s for RNA. During the sequencing process, the double strand is split in front of the pore and only one strand is read. The ionic current per pore serves as a proxy signal for the molecule passing through, is recorded and finally translated into a nucleotide sequence by the basecalling algorithm (Fig. 1.4, right box). Without amplification of the input, the nanopore is sensitive to different base modifications including 5-methylcytosine, 6-methyladenine and even synthetic base analogues. Among other devices distributed by Oxford Nanopore Technologies, the MinION in particular is gaining prominence. The portability and very low acquisition costs open new perspectives of real-time sequencing in the field or in clinical settings.

1.4 Results

The submitted work is structured into four major parts and moving from a zoomed out view into the literature to the development of a universal nanopore data processing pipeline. Further on, a set of nanopore signal processing methods forms the baseline for the subsequent application development for human genetics. This section provides a brief overview of the contribution of this work to the field of long-read nanopore sequencing.

Chapter 2 - State of the Art

Entitled as *the third revolution in sequencing technology*, Van Dijk et al. [16] outline the potentials within an innovative and rapidly developing field of research using third generation sequencing. It is indispensable to keep track of the latest developments in the nanopore field, for instance highlighted by the single read accuracy improvements from 87% to 95% modal, solely by enhanced basecalling algorithms applicable to existing data. This chapter aims to provide a comprehensive overview of the use cases for nanopore sequencing, backed by the computational analysis of millions of publications forming a literature graph of keywords and citations. Lastly, based on in-house data, throughput and accuracy summaries for sequence and methylation detection complete the assessment of the current status of the technology.

Chapter 3 - Nanopype

The widespread availability of the nanopore sequencing platform is sparking the development of novel bioinformatics software. Nonetheless, the streamlined raw data handling and processing of more complex workflows lacked consistent pipelines,

impeding the reproducibility across projects and labs. This chapter covers the development of the modular *Snakemake* pipeline *Nanopype*. A set of nanopore specific workflows cover the most common use cases such as basecalling, alignment, methylation and structural variant detection. Deployed as python module with automatically build and tested software containers, the pipeline maintains all of its internal dependencies, including the version control of integrated tools. *Nanopype* is the baseline for subsequent projects and facilitates the rapid development of custom and integration of third party software. *Nanopype* is published in Bioinformatics: *Nanopype: a modular and scalable nanopore data processing pipeline*.

Chapter 4 - Signal Processing

The interplay of a protein nanopore and molecules passing through it causes a characteristic ionic current signal. In the past, advancing algorithms led to improved single read accuracy, detection of epigenetic base modifications and efficient barcode de-multiplexing based on raw nanopore reads. For the development of novel applications, the raw signal may be the input of choice, allowing to bypass basecaller induced errors. This chapter covers basic raw signal processing methods including simulation, normalization and signal alignment. A flexible raw signal framework enables the seamless integration of signal and sequence space information. The signal processing methods form the algorithmic basis for the repeat quantification method developed in the next chapter.

Chapter 5 - STRique

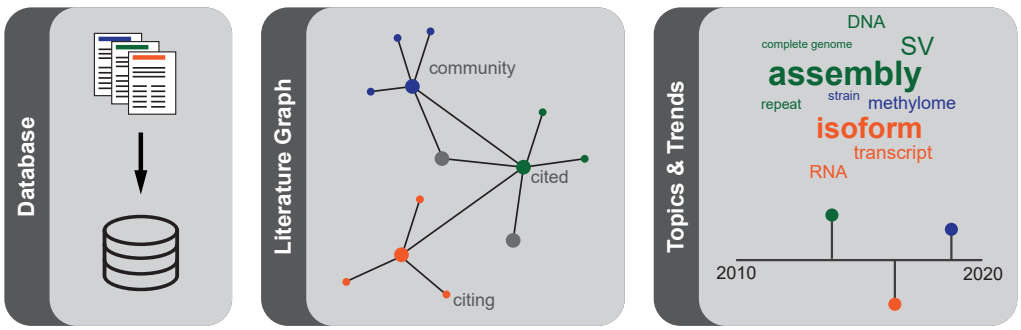
Sequencing of genomic regions, previously inaccessible for short read technologies, is a key advantage of long reads. Stretches of repetitive DNA with low sequence complexity are difficult to investigate once they become longer than the read length. Short tandem repeats (STR) are an example of repeat elements being expanded to multiple thousand nucleotides in length in disease cases. Enabled by our processing framework and signal analysis methods, *STRique* facilitates the precise analysis of STRs in synthetic and patient samples. *STRique* solves the problem of counting repeats on individual read-level based on noisy long-read sequencing data. Our method allows for the first time to exactly quantify the length of STRs and integrate repeat count and methylation state on a single-molecule level. *STRique* is published in Nature Biotechnology: *Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing*.

This thesis concludes with a summary and discussion of the results. An outlook includes the assessment, where third generation sequencing technologies in general and nanopore sequencing in particular are presumably replacing short reads, where they are appropriate to supplement and where short reads are likely to stay state of the art. Taken together, this work aims to provide an overview of the nanopore field in general, propose streamlined processing and novel signal analysis methods and contrast opportunities against challenges within a technology driven research branch.

2

State of the Art

Single-molecule real-time- and nanopore sequencing are commonly referred to as third generation sequencing technologies. Continuous improvements on platform, library preparation and analysis software still lead to throughput and accuracy enhancements. Following latest developments in third generation sequencing is therefore equally important for both, users and developers. The availability of significantly longer reads enables novel insights, published in a rapidly growing number of studies, making a systematic and unbiased manual literature research increasingly complex. The following review is therefore backed by the computational evaluation of meta data from publications in scientific journals. Combining title, abstract and citations into a literature graph opens a unique perspective and provides a scalable approach to sweep any number of publications. In addition to the major application fields of assembly, structural variant and isoform detection, we find a largely separated landscape, either relying on Pacific Biosciences or Oxford Nanopore Technologies. Supplementary code for this chapter is available at *https://github.com/giesselmann/scholar*.



The chapter starts with a **background** in 2.1 followed by the setup and usage of a **literature database** containing scientific publication meta data in 2.2. A big picture overview of **third generation sequencing** technologies in 2.3, is followed by a focus on **nanopore sequencing** in 2.4. Finally, most recent **throughput and accuracy** benchmarks on in-house data close the state-of-the-art evaluation.

2.1 Background

The number of studies published per year in scientific journals is exponentially growing (Fig. 2.1). Including only records with a digital object identifier (DOI) tracked by CrossRef and Semantic Scholar, results in a conservative estimation of 100k journals and a total of 5M paper being published only in 2020. While the targeted discovery of specific studies remains feasible by indexing in search engines, the extensive and continuous tracking of an entire field of research becomes increasingly difficult. Especially for the fast evolving field of nanopore sequencing, a zoomed out perspective is equally valuable for the orientation of newcomers and adaptation to latest developments in experienced groups. A data driven literature scan facilitates the clustering of results and identification of key publications in an unbiased way.

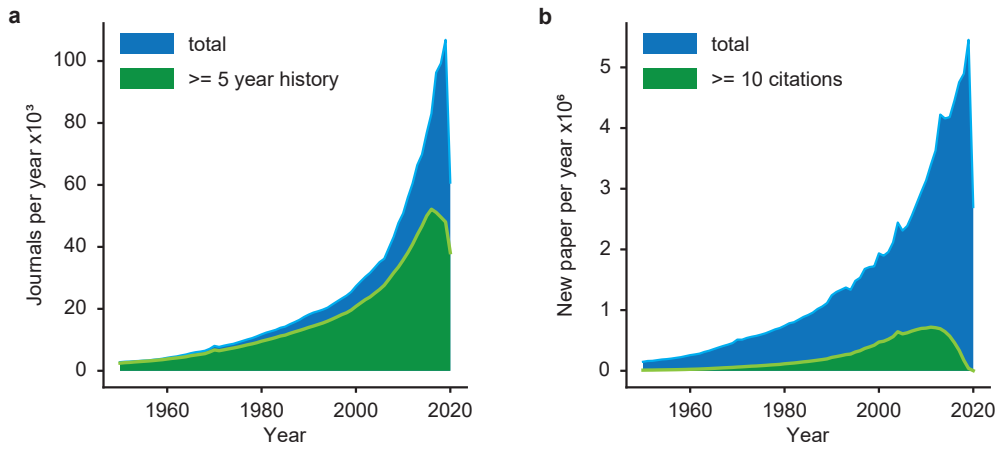


Fig. 2.1.: Journals and new publications per year: **a**, Actively publishing scientific journals per year and journals with a publication history of at least five years. **b**, New publications per year and publications with at least 10 citations. (Semantic Scholar and CrossRef combined, only records with DOI)

The following computer aided literature analysis is inspired by the Open Syllabus Galaxy[1] project. Based on co-assignments of books in North American university courses, their literature graph of 160k books visualizes the linkage between research fields. Applied to scientific journal publications, the success of the method is limited by the availability and quality of large scale meta data such as title, abstract and citations. A number of platforms like Google Scholar, Web of Science, Dimensions or Microsoft Academic operate online literature databases, though without access to larger data chunks for systematic offline analysis. Additional full text for advanced text mining is commonly only available through paid access from individual journals.

[1] `https://galaxy.opensyllabus.org/`, accessed 01/2021

2.2 Literature Database

For the purpose of tracing citations and clustering larger numbers of publications, we first setup a custom literature database. The Semantic Scholar open research corpus (S2ORC) provides the largest available collection of scientific paper meta data and serves as starting point in this work [18]. Collected 07/2020, the data contains 77M papers linked by 333M citations. Provided as compressed JSON files, the records are re-organized into a SQLite database with two main tables for records and citations. Each record is uniquely identified by it's DOI and can additionally contain year, journal, title and abstract. Citations are unique pairs of citing and cited DOI, referencing rows in the records table. To further improve completeness of records and citations, we incrementally query the CrossRef REST API[2] for novel entries, most recently in January 2021. Citations are part of CrossRef but not provided through the API. The CrossRef Open Citations Index (COCI) however, is regularly parsing and dumping citations and therefore integrated as well [19]. Metrics of the final database used for this work are summarized in table 2.1. For comparison, the Dimensions[3] online platform lists 114M records and 1.3G citations.

Tab. 2.1.: Literature database metrics

citations (edges)	901 M
paper (nodes)	116 M
with title	115 M
with title & abstract	57 M
connected nodes	68.4 M
> 0 citations	56.6 M
> 5 citations	27.5 M
largest connected component	67.9 M

The network of publications and citations can also be interpreted as a literature graph with papers as nodes and citations as edges. Only a fraction of 68M publications has either citing or cited edges. The lack of citing (outgoing) edges indicates insufficient parsing of the papers reference sections, while the lack of cited (incoming) edges is likely a combination of technical limitations and unrecognized publications. Striking though, is the presence of the largest connected component of 67.9M publications, where each paper can be reached via at least one citation edge.

A subset of publications originating from the S2ORC data set is annotated with a primary field of research from Microsoft Academic (MA). To get a first impression of the value of a literature graph, we first extracted all edges connecting papers with an

[2] `https://github.com/CrossRef/rest-api-doc`, accessed 01/2021
[3] `https://www.dimensions.ai/`, accessed 01/2021

annotated field of research. Next, a hierarchical clustering of summarized citation counts per field is plotted as heatmap in Fig. 2.2, and shows an intuitive grouping of natural- and social sciences.

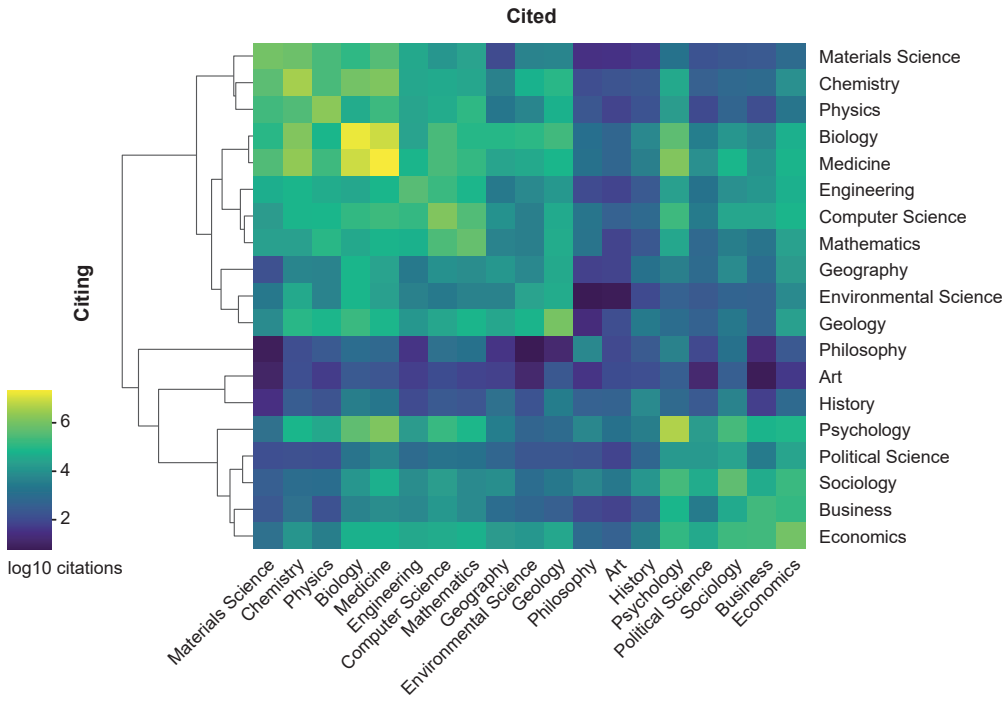


Fig. 2.2.: Citation edges summarized by field of research: Sum of citations (log10) between and within fields of research (Hierarchical clustering, method: centroid, distance: cosine).

A graph is a powerful data structure, for instance to detect local communities or compute distances between nodes, however, with a growing number of nodes and edges it becomes impossible to visualize. For the purpose of visualization and clustering, a graph embedding, similar to the *node2vec* algorithm used in the OpenSyllabus project is therefore needed [20]. A graph embedding is a fixed length vector representation of each node, with the aim to preserve local connectivity.

The vector representation is enabling distance based clustering methods such as KMeans and dimensional reduction by principal component analysis (PCA) and uniform manifold approximation (UMAP). Additionally, while sub-sampling nodes of a graph would split connected components, sub-sampling from an embedded graph, based on e.g. citation counts or topic, preserves the overall structure and reduces required compute resources.

Embedding the scientific literature graph using *node2vec* is due to run-times of multiple weeks on CPU and large memory requirements on GPU not feasible. We use therefore the *DeepWalk* [21] algorithm with its GPU implementation in *GraphVite* [22]. In order to fit on a GPU-server with 4x NVIDIA 2080Ti (11 GB RAM), only

edges connecting publications with at least 20 citations are considered. The resulting sub-graph, in the following referred to as **core_20** graph, contains 9M nodes encompassed by virtually a single connected component.

Embedding with default parameters yields feature vectors of length 128 for each publication. Following the OpenSyllabus projects workflow, the high-level visualization in Fig. 2.3 is generated by first reducing to 64 dimensions using PCA (85% explained variance), followed by further reduction to two dimensions using a UMAP. For a better overview, only publications with fields considered relevant for this work are shown.

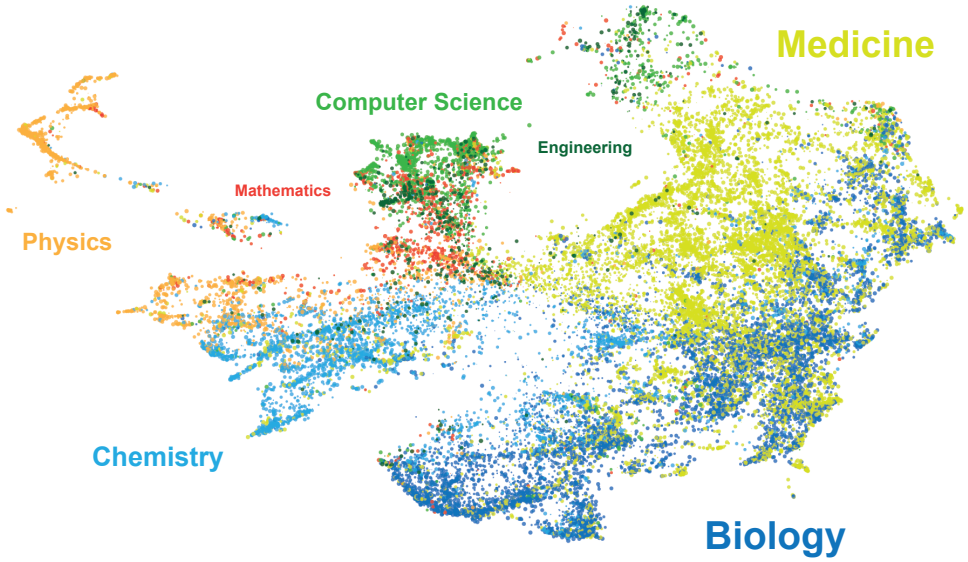


Fig. 2.3.: Scientific publication graph embedding: UMAP visualization of *DeepWalk* graph embedding colored by field of research. Publications from 1980 onward with at least 20 citations (n=9M embedded, n_neighbors 75, min_dist 0.01, metric correlation, n=50k random sample of relevant fields plotted).

In summary, the gathered SQLite literature database supports a variety of fast queries for e.g. research fields, journals, year of publication or citations. The graph embedding of highly cited publications is a proof of concept to visualize literature based on citations and provides an intuitive perspective on field interactions. Due to automated parsing, lack of visibility but also novelty of a paper, a considerable part of the database can not be embedded. To address this issue, a future version with weighted pseudo-citations based on text similarities is planned.

2.3 Third Generation Sequencing

Both, the *Nanopype* pipeline in chapter 3 and *STRique* repeat detection in chapter 5 are exclusively developed for nanopore sequencing data. Nevertheless, from a review perspective, both long read platforms are of interest, in particular when portability or initial acquisition cost play a secondary role. The identification of third generation sequencing related publications is implemented in three steps of seed, extend and connect. First, publications with keywords in either title or abstract are marked as seed paper. The second step extends clusters around seed papers by following their citation edges in the literature graph. Lastly, additional edges derived from text similarities and random walks in the **core_20** backbone graph reduce the sparsity of the third generation citation graph.

Title and abstract columns of the records table are indexed for full text search using the SQLite *FTS5* extension. For PacBio SMRT sequencing, the respective query is 'pacbio OR single molecule real time sequencing', for ONT 'nanopore sequencing' is used. *FTS5* is case-insensitive and works based on tokens, 'sequencing' for instance, is stemmed to 'sequenc', requiring a post processing step to validate matches and reduce false positives. The number of seed papers for PacBio is 2900, for ONT 2852 paper contain the words 'nanopore' and 'sequencing' in title or abstract. The output is comparable to the online platform Dimensions, which lists 3147 publications for 'nanopore sequencing' (queried 01/2021).

The extension of paper clusters around keywords is achieved by iteratively adding publications with a minimum number of citations and with a strong binding to the cluster. The binding is computed as the fraction of in- or outgoing edges reaching nodes in the current cluster. The seed extension is evaluated for different parameter combinations, with five citations and a minimum binding of 0.2 resulting in the largest stable cluster (Fig. 2.4). The PacBio and Nanopore cluster grow to similar sizes of 4392 and 4689 publications respectively and form the third generation sequencing cluster. The number of nodes reached by both extensions though, is with 308 unexpectedly low.

To further link isolated connected components and for additional preservation of the global graph structure, we complement the extracted third generation sub-graph by random walks in the **core_20** backbone graph. Each walk starts at a sub-graph node which is also contained in the **core_20** graph and follows citation edges within the **core_20** backbone, until another shared node is found, or a maximum length is reached. Edges of random walks are only used for embedding, the additional

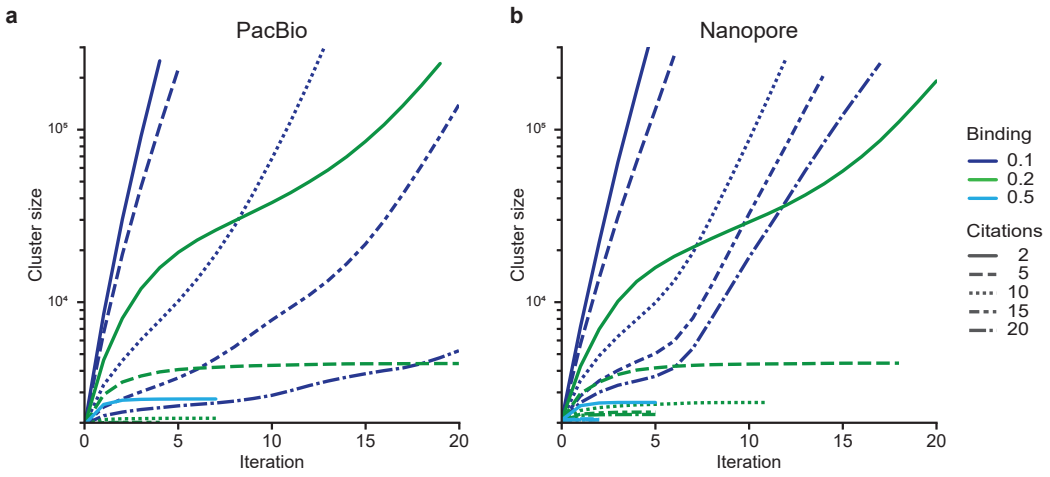


Fig. 2.4.: Cluster size convergence of keyword seed and extend strategy for PacBio query (**a**) and ONT query (**b**). The iteration is stopped on convergence, after 20 steps or on exceeding a threshold of 50k publications.

nodes are ignored in the subsequent content analysis. Of the initial 8.7k third generation sequencing sub-graph nodes, only 7.6k have citing or cited edges suitable for embedding. In order to also integrate publications without citations, we add additional weighted edges based on text similarities between documents. Specifically, a MinHash LSH Forest is used to perform top-k queries based on approximated Jaccard similarities [23]. The LSH Forest is build of hashed titles and abstracts of all publications in the random walk expanded third generation sequencing graph. For each publication, edges to the top 500 most similar documents in the set are stored as candidates, with the approximated Jaccard similarity as weight. Of all candidate edges, the top 20% edges between highly similar documents are added to the third generation sequencing graph. For the embedding, the graph is treated as un-directed and weighted, with a constant weight of 1.0 assigned to all citation edges. Applying the same *DeepWalk* configuration and dimensionality reduction as above results in the third generation sequencing graph embedding shown in Fig. 2.5.

The third generation literature graph embedding reveals a largely separated field of publications relying on either Pacific Biosciences or Oxford Nanopore Technologies. Aside from small outlier groups, a broad PacBio field is accompanied by three distinct accumulations of nanopore related work. The embedding of the third generation sequencing sub-graph is clustered after discarding nodes added by the random walk expansion. Using the KMeans algorithm, an optimal number of clusters is found with 24 based on the distortion score. A subsequent hierarchical clustering of cluster centers results in a distance based sorting (Fig. 2.6 a). The ratio of seed to extension papers per technology and cluster confirms the previously noted separation of PacBio and ONT (Fig. 2.6 b). Aside from small clusters of virtually only extension

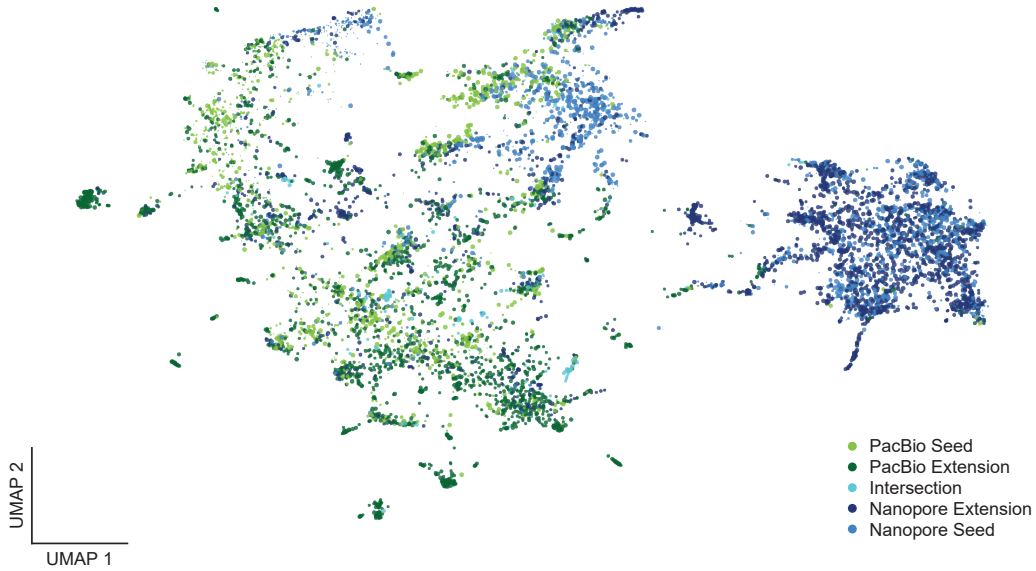


Fig. 2.5.: Third generation sequencing cluster: The UMAP projection shows the *DeepWalk* graph embedding with points scaled by number of citations. Colors indicate technology and seed/extend source.

publications (ID 0, 6-7, 17-19), two major groups, of which one (ID 20-23) contains only nanopore seeds and extensions, become visible. Remarkable is, that clusters with on average older publications tend to have a higher proportion of extensions, indicating, that the extension of the novel third generation field is primarily directed towards older publications, defining the broader field of application. In return, the clusters with the highest mean citation counts are 12, 15, 22 and 23, each with high seed to extend ratios. These clusters are in the following identified as core-assembly, MinION usage, solid state pore development and analysis of pore translocating polymers.

Word clouds provide an intuitive way for high-level content analysis of each cluster. First, titles and abstracts are split into word tokens and further reduced to their stem using the Porter Stemming Algorithm. Descriptive tokens are extracted using a term-frequency inverse-document-frequency (TF-IDF) embedding. After filtering out tokens occurring in less than 10 documents or more than 20% of the corpus, each token in each document is assigned a score based on how frequent it is mentioned in the document (for titles and abstracts commonly once) as opposed to how frequent it is found in the corpus. The result is a word embedding matrix of 8k third generation sequencing publications times 4.6k descriptive word tokens. After averaging TF-IDF scores per cluster, each clusters top 200 highest scoring tokens are illustrated as word clouds, of which three examples are shown in Fig. 2.7.

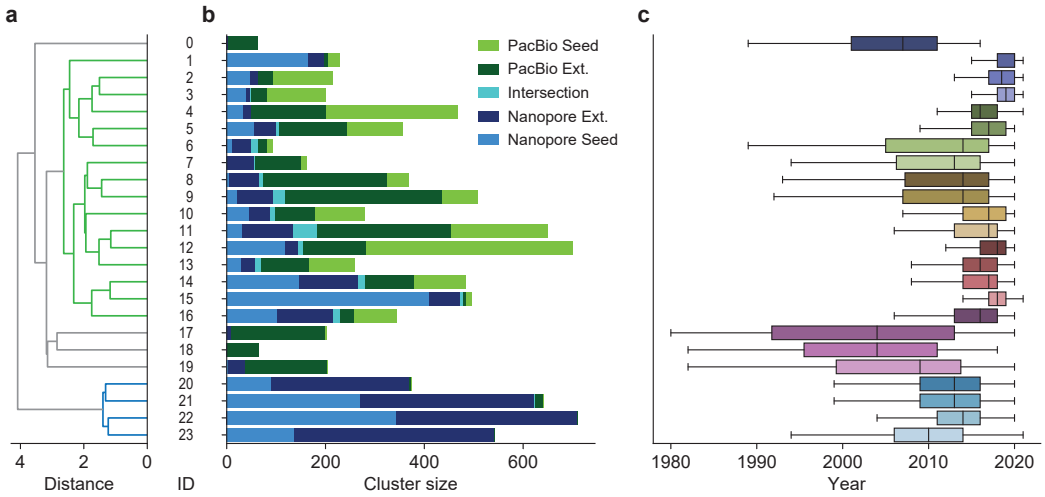


Fig. 2.6.: Cluster of long-read applications: **a**, Hierarchical clustering of cluster centers assigns cluster IDs based on inter-cluster distance. **b**, Cluster size, seed to extend and technology composition are shown as stacked bar plots. **c**, Distributions of publication years per cluster with data as boxplots (centerline, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range).

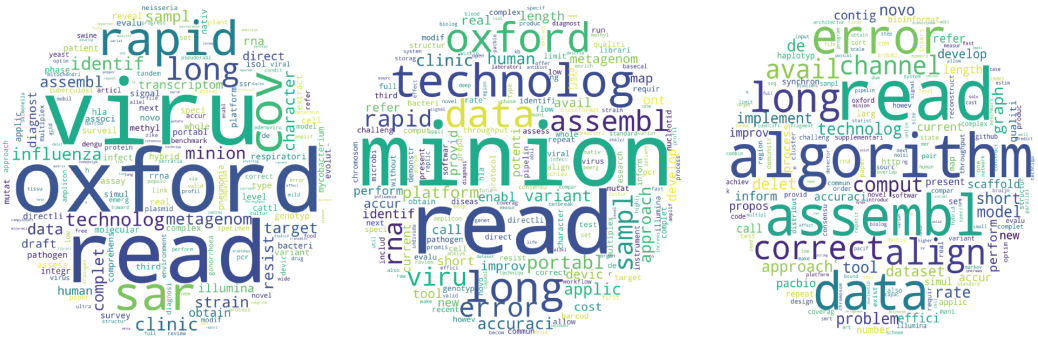


Fig. 2.7.: Nanopore sequencing word clouds of cluster 1, 15 and 16 reveal a virus identification, a composite cluster around the usage of the MinION and an algorithmic cluster around assembling, error correction and alignments.

Not all word clouds can be translated into meaningful cluster labels. Especially cluster with few publications or a low seed to extension ratio are difficult to characterize. Nonetheless, for most clusters a high-level label is found and annotated into the graph embedding (Fig. 2.8). Surprisingly, while referring to nanopore sequencing, the isolated clusters (ID 20-23) have no direct reference to Oxford Nanopore Technologies, but cover the development of novel protein and graphen based nanopores, translocating polymer and signal analysis. Beyond that, the token 'assembl' (words: assembly, assemble, assembling etc.) is one of the most prominent, and with different scalings found in almost all clusters.

The following section aims to provide a brief insight into major application fields for long-read sequencing technologies, while in particular highlighting the contribution

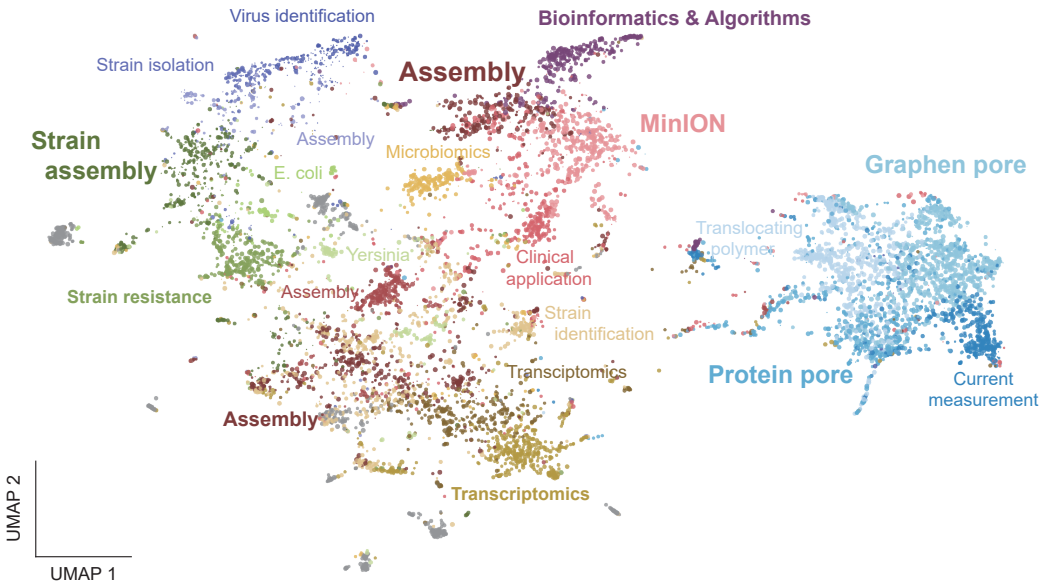


Fig. 2.8.: Third generation sequencing applications: The UMAP projection shows the *DeepWalk* graph embedding with points scaled by number of citations (1 for content linked publications). Colors indicate cluster with labels derived of most descriptive word tokens.

of nanopore sequencing. Key publications are determined as highly cited paper per year, in comparison to their respective cluster community. In clusters with only few seed publications, these are primarily mentioned to emphasize the role of the technology within the field.

Limitations: While so far providing a data-driven high-level overview of the third generation sequencing field, the following section is referring to individual publications based on citation counts. By citing what is already highly recognized, a bias against very new or comparably small application fields is anticipated. An example are the already visible tokens 'sar' and 'cov' in the virus identification cluster. Contributing to an emerging application field of assembling variants and detecting mutations of the SARS-CoV-2 virus, associated key publications based on citation counts are still impossible to obtain using our database.

2.4 Nanopore Sequencing Applications

Publications of the composite 'MinION' cluster (ID 15) receive, with an average of 28, the most citations. Followed by 18 for the 'Assembly' cluster (ID 12) and 16 for 'Bioinformatics & Algorithms' (ID 16), these clusters form the most recognized section of the third generation sequencing field. None of the other clusters is reaching more than 11 citations on average. Based on a first inspection, these highly cited clusters

contribute primarily methods and proof of concept studies, while the surrounding clusters define the broader field of application with in detail very specific studies. Major application fields are identified as:

1. Bioinformatics
2. Assembly
3. Structural variant detection
4. Bacterial and viral strain analysis
5. Metagenomics and microbiomics
6. Isoform detection

For the 'MinION' cluster in general and otherwise where appropriate, publications are mentioned in the context of their primary usage, regardless of their cluster assignment. An interesting example is the interplay of 'Bioinformatics & Algorithms' and the 'Assembly' arm of the bacterial strain analysis field (Fig. 2.6 ID 16 and 3, Fig. 2.8 light purple and dark red). While a variety of tools and pipelines for *de-novo* genome assembly have been published, the most prominent *Canu* [24] is assigned to cluster ID 3, pulled towards its application to bacterial and viral strains. A central cluster with respect to the UMAP projection (Fig. 2.8, ID 14) covers clinical applications. Yet individual publications rely on methods of the surrounding clusters, for example detection of disease associated structural variations or analysis of antibiotic-resistant strains. So far not visible as independent cluster, but of particular interest for this work, is the detection of epigenetic base modifications using nanopore sequencing. An according section with current methods and proof of concept studies is therefore appended.

2.4.1 Bioinformatics

Significantly longer reads, in combination with higher error rates compared to NGS technologies, require adaptation of existing and development of novel bioinformatics software. Among the earliest publications are tools providing **simulations** of SMRT sequencing data [25, 26]. Virtually unlimited amounts of artificial long reads with characteristic length distributions and error profiles facilitate early development and testing of downstream applications, independent of the real data throughput at the time. Similar tools exist for nanopore signal and sequence data, but are not visible by their citations. With existing data and experience from SMRT sequencing, we hypothesize, that the nanopore versions are primarily for the development and

debugging of for instance streaming algorithms and become less important with more real data being publicly available.

Despite ongoing improvements, the error rate of third generation sequencing technologies requires constant quality control. Particularly for nanopore sequencing, a number of **visualization** packages have been developed. Initially receiving a lot of attention, none of *Poretools* [27], *PoRe* [28] or *NanoOK* [29] is actively maintained and would be able to handle the most recent ONT file formats. Functional and under active development is *NanoPack* [30], supporting both, SMRT and nanopore data.

Translating the primary signal into sequences during **basecalling** is a crucial processing step for third generation sequencing reads. High-fidelity (HiFi) SMRT reads and advancing algorithms for nanopore basecalling render initial read-level error correction methods obsolete [31–33]. For nanopore reads, the single read accuracy has increased from ~85% using the cloud-based *Metrichor* over ~87% in *albacore* to 95% in the most recent recurrent neural network basecaller *guppy*. Clearly headed by the manufacturers, a couple of community methods have been proposed, namely *DeepNano* [34] *Nanocall* [35] and *Chiron* [36]. Although free and open-source, these are not competitive in terms of accuracy and run time. The performance of neural network basecalling tools for ONT data has recently been reviewed [37]. Yet, the ongoing improvement of production grade, and promising experimental platforms such as *bonito*[4], require constant quality assessment on patch to patch level.

Shared by most downstream applications is the **alignment** of reads against a reference genome. The first recognized long read aligner is *BLASR* [38], maintained by PacBio and for SMRT reads only. Published in 2016, *minimap* [39] and *GraphMap* [40] are both developed for long, noisy read alignments. While *minimap* computes very fast approximate mappings, *GraphMap* is very sensitive, but requires, due to its reference index structure, considerably more memory. Actively improved and maintained in version two, *minimap2* [41] and *GraphMap2* [42] additionally support spliced alignments of transcript reads. The fast and memory efficient algorithm of *minimap2* makes it to one of the most used aligners in the field. An aligner, especially developed for the mapping of structural variations is *NGMLR* [43]. Splitting of reads into fragments of few kilobases, and subsequent individual alignment using a convex gap cost model, improves the accurate detection of breakpoints. Using parameter presets, *minimap2*, *GraphMap2* and *NGMLR* are capable of mapping both, SMRT and nanopore reads.

[4] `https://github.com/nanoporetech/bonito`

2.4.2 Assembly

De-novo assembling of previously unknown and improving the quality of existing draft genomes is a primary application of long read sequencing. Assembling greatly benefits from increased throughput per flow cell, read-lengths and single-read accuracy, replacing next- by third-generation technologies as current gold standard. Both, PacBio and ONT enable state-of-the-art genome assemblies, especially for viral and bacterial genomes. The choice of a suitable platform therefore also depends on already available data, portability and instrument acquisition cost. Limited by throughput and accuracy, initial SMRT assemblies focused on the scaffolding and completion of existing bacterial draft genomes [44, 45]. Hybrid approaches combine next- and third-generation sequencing data, and thus integrate the accuracy of short reads with the continuity of noisy long-reads [46, 47]. First third-generation-only assemblies include a finished fungal genome (*Verticillium dahliae*, 35Mb) [48] based on SMRT reads and optical mappings and a nanopore-only assembly of *Escherichia coli* (4.6Mb) [49]. More recently, studies report the successful assembly of ~80% of the hexaploid wheat genome (*Triticum aestivum*, 17Gb) with NG50s of 88kb and 480kb [50, 51] and an improved maize reference (*Zea mays* , 2.1 Gb, NG50 1.2Mb) [52], each using SMRT sequencing in combination with optical mapping. Further examples of improved genome sequences are tea (*Camellia sinensis var. sinensis*, 3.1Gb, NG50 1.4Mb) [53] and apple (*Malus x domestica*, ~700Mb, NG50 7.0Mb) [54], with the latter using a combination of optical mapping and Hi-C (Chromosome conformation capture) data for scaffolding. In summary current state-of-the-art chromosome-scale assemblies of large, potentially di- or polyploid genomes can be build solely based on SMRT or nanopore reads in combination with optical mapping and/or Hi-C data for scaffolding [55].

A number of standalone tools and pipelines for genome assembly have been published. Based on their citations, *Canu* [24] and *Unicycler* [56] are the most used assembly pipelines. Both are not assigned to the core-assembly cluster (ID 12, 13), but are part of the strain-assembly and strain-resistance cluster respectively, highlighting their importance for a large user group. Developed exclusively for SMRT sequencing is *FALCON* [57], with the additional function of phasing diploid genomes. *Racon* [58] and *miniasm* [39] are lightweight standalone assembly tools, both assigned to the 'MinION' composite cluster. Together they are forming the computational backend of *Unicycler* when processing long-reads and *Racon* is performing the initial consensus assembly of transcript reads in the ONT isoform detection pipeline *Pinfish*[5]. Two important standalone assemblers published in 2020 are *Flye*

[5] `https://github.com/nanoporetech/pinfish`, accessed 02/2021

[59] and *Wtdbg2* [60], both missed by the data driven state-of-the-art analysis so far. Both abstracts do not contain a direct reference to either one of the long-read technologies and despite having already high citation counts, both are not reached by the extension strategy. This example points to two shortcomings of the proposed literature mining strategy: First, it is only sensitive to topics (tokens) present in title and abstract. Second, its success and coverage are decisively impacted by the seed-extension strategy. Nonetheless, if specifically interested in assembling, the token 'assembl' would have picked up both examples as seed publications.

Similar to the continuous improvement of basecalling algorithms, the assembly field is advancing at a fast paste. Published in 2018, Jain et al. report the linear assembly of the centromere on the human Y chromosome and later the assembly of a human genome using only nanopore long-reads [61, 62]. Only two years after publication, the reported NG50 of 6.4Mb based on *Metrichor* basecalling and *Canu* assembly is clearly outperformed by more recent workflows. An in-house benchmark of *Flye* on a human data set with 30X coverage and 20kb N50 read length lead to a draft genome assembly with 25Mb NG50, further outperformed by the developer benchmark, reporting up to 40Mb NG50 using comparable coverage[6].

2.4.3 Structural Variant and Haplotype Resolution

Structural variations (SVs) describe differences, for instance deletions, insertions, or translocations, between individual genomes and a reference sequence. On individual base level, single-nucleotide polymorphisms (SNP) are substitutions of single bases in the genome, observable in a fraction of a population. Both variation types are detectable by NGS technologies, advantages of long-read sequencing are the accurate detection of large variants and the long-distance haplotype resolution based on SNPs. With no dedicated cluster for structural variant detection, respective tools got assigned to the bioinformatic cluster (ID 16), while applications are mostly found within the clinical application cluster (ID 14). Currently available SV detection tools include *NanoSV* [63], *Sniffles* [43] and *SVIM* [64]. With sequence alignments as inputs, these are commonly applicable to SMRT and nanopore reads. The performance is, however, influenced by the preceding alignment, finding the combination of *minimap2* and *Sniffles* the currently most sensitive and fastest option [65].

The analysis of human leukocyte antigen (HLA) variants is a major clinical application relying on both, NGS and third-generation sequencing. The polymorphic HLA genes are essential for the human immune system. Before organ transplantation,

[6] `https://github.com/fenderglass/Flye`, accessed 02/2021

their characterization is essential to minimize the chance of rejection by the receiving patient. In addition to high-throughput NGS methods, both SMRT and nanopore sequencing are applied to detect HLA variants [66, 67], leading to characterizations of novel and confirmation of know alleles [68]. Furthermore, long-read sequencing is used to characterize disease-associated structural variations, for instance segmental duplications in patients with Potocki–Lupski syndrome (PTLS) [69] or deletions causing tumor suppressor gene inactivation in pancreatic cancer [70]. Nonetheless, NGS still plays a major role, especially in the analysis of large patient cohorts, where available library preparation input material and sequencing cost have to be taken into account [71].

2.4.4 Bacterial & Viral Strain Analysis

The identification and analysis of viral and bacterial strains is, in terms of publication counts, the largest application field for long-read sequencing. Significantly smaller genomes compared to mammals and plants facilitate the early adaptation of third-generation technologies by compensating the initial error rates with high sequencing coverage. In stationary operation, the PacBio Sequel II and the ONT MinION appear to be equally well suited platforms. However, based on citation counts, publications relying on the real-time and mobility aspect of the MinION stick out in the following. In contrast to the massively parallel SMRT sequencing, individual reads from a nanopore sequencer can be analyzed directly after passing through the pore. Proof of concept studies have demonstrated mobile nanopore sequencing in challenging environments on the International Space Station [72] and in the Antarctic [73]. This section is subdivided into the mobile identification of viral strains (Cluster ID 1), analysis of bacterial antibiotic resistance (ID 5, 6) and metagenomic identification of bacterial and viral pathogens (ID 15).

The ability to perform in-field sequencing is a unique feature of the ONT MinION device. The 2015 outbreak of Ebola in West Africa [74], the Zika virus epidemic 2016 in Brazil [75] and the 2018 Lassa fever outbreak in Nigeria [76] have been monitored using nanopore sequencing, proving the feasibility of on-site metagenomic analysis. The studies demonstrate direct sequencing of viral genomes without the need for complex laboratory infrastructure. Potential applications include guidance of control measures by reconstructing epidemic origins, monitoring transmission rates of different strains and generation of genomic information for vaccine development.

"*Preventing, reducing, and controlling the emergence of antimicrobial-resistant organisms is a major public health challenge requiring the participation of the entire*

medical community and public health agencies." [77] Nanopore sequencing is used to identify bacterial antibiotic resistance genes [78–80]. In a clinical environment the sequencing, assembly and antimicrobial resistance gene annotation has been demonstrated in less than 6h [81].

Furthermore, the applications of nanopore sequencing in a clinical setting includes the identification and characterization of viral [82] and bacterial [83] pathogens. An example is the same-day antibiotic susceptibility prediction of Mycobacterium tuberculosis with a turnaround time of 7.5h [84] Recent throughput advancements enable multiplexed sequencing of multiple samples on a single flow cell and thus improve the economic efficiency [85].

2.4.5 Microbiomics

An isolated cluster (ID 10, Fig 2.8 orange) of comparable high density covers long-read sequencing in micobiomics. Two major identified applications are the improved assembly of bacterial genomes from microbiomes [86, 87] and phylogenetic profiling using 16S rRNA sequencing. The 16S ribosomal RNA is a conserved subunit of prokaryotic ribosomes, allowing the species identification in complex microbiomes. The longer reads of SMRT sequencing have become a cost efficient alternative to Roche's 454 Genome Sequencer in studying microbial diversity [88, 89]. Based on publication counts, PacBio appears to be the primarily used platform, however, the portability and decreasing error rates of the ONT MinION make it an alternative to consider [90].

2.4.6 Isoform Detection

Both third-generation platforms support the sequencing of full-length RNA transcripts after conversion to cDNA. In addition are the nanopore devices by ONT capable of directly sequencing RNA, providing the potential to analyze RNA modifications. For RNA sequencing, the number of reads is crucial in order to capture a maximum amount of transcripts. The parallel SMRT sequencing is capped at around 4M reads per flow cell, whereas the throughput of a nanopore cDNA run is impacted by flow cell and library quality, ranging from 5 to 10M reads. The direct RNA nanopore sequencing is performed at lower speed with only 70nt/s (450nt/s DNA), resulting in lower throughput of around 1M reads per MinION flow cell.

SMRT and nanopore sequencing are able to characterize and quantify transcripts, with improved accuracy by employing short-read hybrid approaches [91, 92]. However, ambiguous alignments and frequently truncated fragments make isoform detection and quantification more challenging than anticipated [93]. Applications utilizing the full-length transcript sequencing include characterization of novel transcripts [94], expression analysis of human LINE1 transposable elements [95] and estimation of poly(A) tail lengths [96].

Frequently mentioned in the context of nanopore direct RNA sequencing is the ability to detect RNA modifications. Current methods to detect m6A in nanopore transcript reads are based on the error profile of the used basecalling algorithm [97, 98]. Highly dependent on the basecaller version, these approaches can only be proof of concepts, demanding for more reliable signal based methods.

2.4.7 Base Modification Detection

Single-molecule real-time sequencing is sensitive for a number of base modifications, in particular N6-methyladenine (6mA) [99]. Modifications impact the kinetics of the polymerase while synthesizing the complementary strand, leading to characteristic pulse duration changes. The sensitivity of SMRT sequencing for different modifications varies, the recommended coverage is ranging from 25X for 6mA to 250X for 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC)[7]. With a focus on mammalian genomes and a primary interest in 5mC, such high coverage is, aside from sequencing costs, also not allowing for any single read analysis. The sensitivity of nanopores to detect chemical base modifications, in particular 5mC and 5hmC, has already been reported in advance to the release of the ONT sequencing platform [100, 101]. Modified bases cause a characteristic level change in the ionic current signal, which can be detected by downstream analysis software. Both third-generation technologies can at individual-read and -base level only report a modification probability, leading to overall lower accuracy and missing values compared to whole genome bisulfite sequencing.

Nanopolish [102] and *SignalAlign* [103] are the first tools detecting base modifications based on reads from the ONT MinION platform. Developed for the R7 pore generation, *Nanopolish* is reporting 5mC, *SignalAlign* 6mA, 5mC and 5hmC. However, only *Nanopolish* is maintained, supporting the R9 pore and current raw signal file

[7] `https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf`, accessed 01/2021

formats. More recently, neural network based approaches in ONT *Megalodon*[8], *DeepMod* [104] and *DeepSignal* [105] show promising accuracy improvements reviewed in [106]. Yet, broad usability is hampered by non-standard output formats, lack of maintenance and complex software dependencies such as the ONT *tombo*[9] signal alignment and the *tensorflow* environment.

Nanopore sequencing in epigenetics appears to be in a proof of concept and development state, with no demarcated publication cluster visible at the moment. An example of its successful application is the rapid classification of brain tumors using nanopore sequencing, combining structural variant, epigenetic profiling and real-time analysis [107]. Furthermore, a recent proof of concept reports the simultaneous profiling of chromatin accessibility and methylation by induced GpC methylation called nanoNOMe sequencing [108]. Finally, for targeted profiling of 5mC and 5hmC, TAPS sequencing combines the accuracy of short-read bisulfite sequencing with long-read phasing information [109]. Both methods are promising developments towards integration of additional information on single-molecule level and improved detection accuracy of DNA methylation.

2.5 Throughput and Accuracy

Flow cell quality, library preparation protocols and bioinformatic analysis software underwent continuous improvements since the first release of the ONT MinION platform. A similar, but delayed, trend is observed for the high-throughput PromethION sequencer. Rapid development requires the constant evaluation of novel and existing workflows, in order to stay up to date with latest advancements in the field. A MinION flow cell can have up to 2048 pores of which, limited by the number of ionic current sensors, at most 512 can sequence at the same time. Pores are multiplexed to sensors in groups of four, the sequencing software *MinKNOW* is controlling the pore selection in order to maximize the sensor occupancy. In 2017, MinION flow cells had an average of 1200 pores, yielding on average 4Gbp in in-house experiments. These numbers increased to around 1400 pores and an average throughput of 7Gbp between 2018 and 2019. Recently, the flow cell quality is mostly stable at 1400 pores, with consistent throughput of at least 8Gbp and up to 25Gbp from single MinION flow cells (1500-1600 pores) using a nuclease flush protocol. The nuclease flush increases the throughput by unblocking clogged pores. After a certain period

[8] `https://github.com/nanoporetech/megalodon`
[9] `https://github.com/nanoporetech/tombo`

of sequencing time, stuck DNA fragments are digested and the flow cell is reloaded with a fresh library (Fig. 2.9 a).

Starting with ~85% accuracy in 2015 [110], subsequent basecaller generations have improved the single read accuracy to 95% median in the most recent *Guppy* high-accuracy model (Fig. 2.9 b). Importantly, any sample sequenced with the same pore version (currently R9.4) can be re-analyzed and benefit from the latest accuracy.

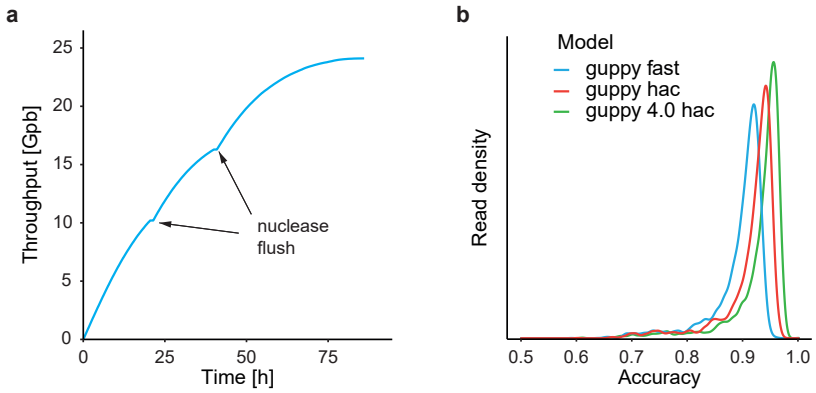


Fig. 2.9.: Throughput and single read accuracy. **a**, Cumulative sequencing throughput over time of one representative MinION flow cell. Nuclease flushes unblock clogged pores and the flow cell can be reloaded with a new library. **b**, Density plot comparing single read accuracy (BLAST identiy) of 4k random reads basecalled with fast and high-accuracy model of *Guppy* v3.5 and high-accuracy model of *Guppy* v4.0 (Median accuracy: 0.90, 0.93, 0.95; Modal: 0.92, 0.94, 0.95).

Without adapter and quality trimming of reads, the throughput of an NGS run equals read count times read length. The read length distribution of a nanopore sequencing run depends on the library preparation, with median and N50 read length as common metrics. The N50 is the minimum read length in the larger half of the total throughput, in other words, half of the total sequencing throughput is supported by reads longer than the N50. Read length distributions, median and N50 are compared for two libraries (Fig. 2.10), of which one included a size selection step. The selection for reads of around 15kB is visible in the distribution, but not reflected in the median (Fig. 2.10 a). The throughput as a function of read length is shown in Fig. 2.10 b, revealing a higher yield (area under curve) for the size-selected library, but a larger N50 for the untreated one. In summary is the restriction of read lengths beneficial for the absolute throughput, however larger N50s are for instance crucial for the continuity of genome assemblies.

Single molecule base modification detection of for instance 5-methylcytosine is a key advantage of nanopore sequencing and of particular interest for this work. Accuracy and scalability of *Nanoplish* are decisive arguments for its use on a daily production

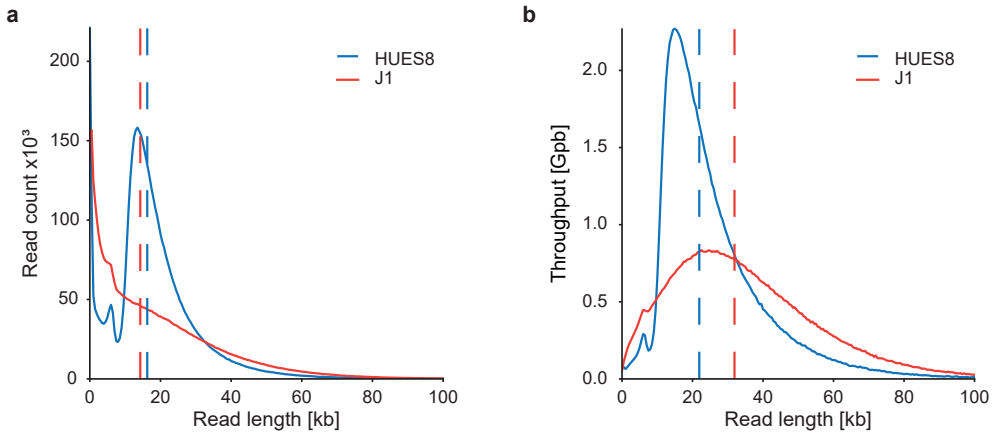


Fig. 2.10.: Read length distribution: **a**, Read length distributions shown as number of reads per length (bins of 500nt) of two PromethION flow cells loaded with libraries from different preparation methods. The preparation of the HUES8 library included an additional size selection. Dashed lines show median read length at 15883 (HUES8) and 14281 (J1). **b**, Read length distributions shown as sequenced basepairs per read length. Dashed lines show N50 at 21041 (HUES8) and 31906 (J1).

level. The accuracy is assessed in comparison to whole genome bisulfite sequencing of the same human cell line in Fig. 2.11 a, resulting in a correlation coefficient of 0.9 on 23M intersected CpGs. The methylation state per CpG on single-read level is reported as a methylation probability (HMM methods) or confidence score (NN methods). Both require subsequent filtering to reduce the overall error rate. The impact of a log-likelihodd threshold on the *Nanopolish* methylation detection accuracy is shown in Fig. 2.11 b. More stringent thresholds reduce the error rate, at the cost of less CpGs passing the filter (Fig. 2.11 c).

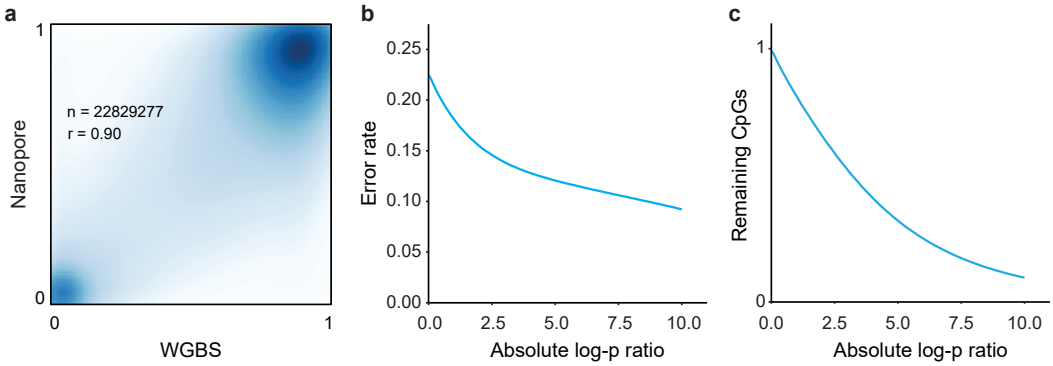


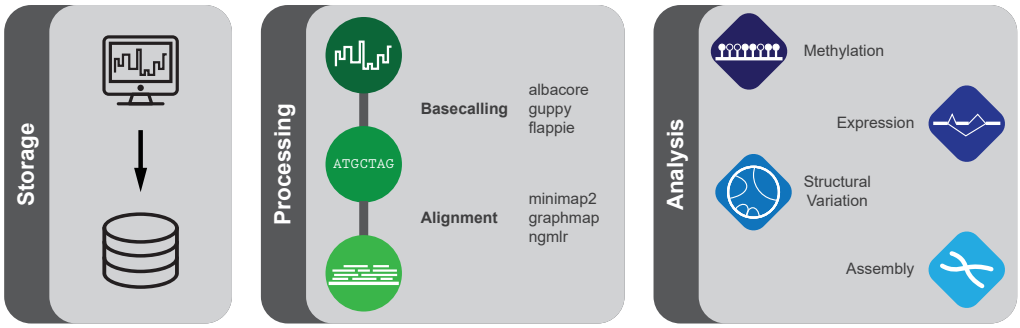
Fig. 2.11.: Nanopore 5mC methylation detection: **a**, Correlation of whole genome bisulfite sequencing (WGBS) and nanopore sequenced mean methylation rates per genomic position (Pearson correlation, *Nanopolish* detection with abs. log-p threshold: 2.5, min. coverage: 10X, reference hg19). **b**, Detection error depending on the applied absolute log-likelihood ratio (methylation probability) threshold. **c**, Fraction of dataset remaining depending on log-p value threshold.

2.6 Summary

In 2015, a year after the release of the ONT MinION, Loman et al. review the potential of nanopore sequencing [111]. Despite the initially high error rates, they anticipate the application in genome assembly, real-time pathogen sequencing in hospitals and environmental monitoring. Five years later, our systematic review of third-generation sequencing literature identifies these topics as major application areas. We demonstrate a data driven literature analysis, suitable to generate a big picture overview of an emerging technology. Citation and content based clustering allows us to characterize application fields of nanopore and SMRT sequencing. Advancements on laboratory and bioinformatic side have promoted long-read sequencing to become state-of-the-art in genome assembly and structural variant detection. On the other hand, persistent shortcomings such as the unchanged high amount (>1ug) of input DNA for library preparation remain and inhibit usage in areas, where input material is rare.

3 Nanopype Processing Pipeline

Long-read third-generation nanopore sequencing enables researchers to now address a range of questions such as genome assembly, structural variant and isoform detection that are difficult to tackle with short read approaches. The rapidly expanding user base and continuously increasing throughput have sparked the development of a growing number of specialized analysis tools. However, streamlined processing of nanopore datasets using reproducible and transparent workflows is still lacking. Therefore we developed *Nanopype*, a nanopore data processing pipeline that integrates a diverse set of established bioinformatics software while maintaining consistent and standardized output formats. Seamless integration into compute cluster environments makes the framework suitable for high-throughput applications. As a result, *Nanopype* facilitates the comparability of nanopore data analysis workflows and thereby should enhance the reproducibility of biological insights. *Nanopype* is available at *https://github.com/giesselmann/nanopype*.



Note: This chapter is based on the publication P. Giesselmann et al. *Nanopype: a modular and scalable nanopore data processing pipeline*, Bioinformatics, 2019 and contains text from the original paper.

The chapter starts with a brief **background** in 3.1 followed by high-level pipeline **design** decisions covering storage, tool encapsulation and reproducibility in section 3.2. Grouped into **modules** individual tools are highlighted in section 3.3. The **installation** section 3.4 illustrates the setup and configuration of the pipeline in different environments. Finally the **usage** on a daily production level is outlined in section 3.5.

3.1 Background

Due to constant development and improvement of applications, frequent reprocessing of the raw signal and downstream data is necessary. Based on the same chemistry and pore version, the single read accuracy could be enhanced from 90% to 95% modal in software over the past years. Thus, novel archiving and processing strategies are needed for data storage and handling that scale with the large amount of data produced by the MinION sequencer. Even when using the most recent compression, the raw signal data is typically five times larger than the sequenced base pairs, resulting in hundreds of gigabytes per sequencing run. This will still be more relevant as higher throughput devices such as the PromethION become more widely available. Furthermore, a limiting factor of the applicability of this new technology are the currently available, research-grade software packages for nanopore long read data analyses. These tend to be difficult to install and require complex software environments. Despite the growing number of recently developed algorithms [112], primary data processing remains challenging due to stand-alone tools without congruent data formats and requirements. Most recent examples of nanopore data processing pipelines are *Katuali*[1] for basecalling and assembly and *Pinfish*[2] for RNA isoform detection from cDNA and direct RNA sequencing experiments. However, they have been developed to perform very specific and inflexible analysis workflows without integrated handling of the critical raw data storage or version control of wrapped tools.

To overcome these issues, we have developed *Nanopype*, a pipeline designed explicitly for streamlined and automated nanopore long read processing. Apart from the integration of essential basecalling, quality control, and alignment tools, we facilitate a set of publicly available analysis applications for barcode demultiplexing, DNA methylation readout, structural variant calling, RNA isoform detection and genome assembly. Based on the *Snakemake* engine [113], our method integrates established error handling and uniform output structures across multiple experiments. Furthermore, *Nanopype* can be run in a parallel setup on both single computers and server clusters. Deployed as a python module, *Nanopype* is mostly built from source with encapsulated routines to simplify the initial setup and integration into existing environments. Additionally, we provide Singularity images for all modules and an automatically built all-in-one Docker container. This enables the usage of the pipeline for both less bioinformatically experienced experimental scientists and bioin-

[1] github.com/nanoporetech/katuali
[2] github.com/nanoporetech/pipeline-pinfish-analysis

formaticians. Lastly, Nanopype provides a well-defined framework for standardized processing independent of the underlying operating system.

3.2 Design

Nanopype's core element is a modular setup to easily update existing tools and to allow seamless integration of the latest developments. Nonetheless, each pipeline release is freezing the included tool versions to guarantee reproducible results. *Snakemake* as workflow management is chosen over the competitor *Nextflow* to support a custom cluster engine and preferring Python over Java for rapid development. In a nutshell, *Nanopype* has been designed around three key components: raw data storage, tool encapsulation and standardized directory structures that mirror the applied toolchain.

Snakemake

Snakemake is a Python based workflow management system. Reproducibility and scaling on cluster environments make it preferable over plain bash script or Makefile based pipelines. As a core concept, *Snakemake* is built around *rules* defining steps to compute an output from an input file. By requesting a single output file, the user initiates the construction of a directed acyclic graph (DAG), chaining all intermediate steps needed to complete the workflow. With the inputs of sequences and reference genome, an alignment rule could for instance compute alignments into an intermediate SAM format to be further sorted and converted to BAM-files. *Snakemake* workflows are defined through the resulting file and folder structure and require a shared file system across compute nodes.

3.2.1 Storage

The first core design component is the consistent storage of the raw signal data from any ONT sequencer. Raw nanopore reads are stored in FAST5 files, an ONT specification of the universal HDF5 file format. Initially, the sequencers exported one file per read, resulting in hundreds of thousands of files per sequencing run. The number of files generated could, especially on Linux file systems, disrupt background services such as nightly mirrors and backups. *Nanopype* is backward compatible with datasets of single read FAST5 files, for which we provide a module to import and package single reads into TAR archive batches. More recently, in 2019 ONT utilized

the full functionality of the HDF5 format with groups (comparable to directories), datasets (structured arrays of primitive data types) and attributes (single values for meta information) and released the multi-read-FAST5 format. The new format typically groups 4000 reads into a single file, resulting in notable improvements regarding copy and synchronization tasks. Lately, the replacement of the *GZIP* by the custom *VBZ*[3] compression reduced file sizes by approximately 30%.

Besides the *Nanopype* pipeline we propose a multi-device and multi-user nanopore sequencing setup (Fig. 3.1). Provided with a suitable server infrastructure, the design supports processing of multiple sequencing runs per week from both local and remotely connected devices. Deploying *syncthing*[4] on device and server side, we synchronize the output folder from the ONT sequencing software *MinKNOW* with device specific folders in a central spooling area. After completion of the sequencing, the data is moved to the active storage, the file ownership is changed to a particular *data user* and files are made write-protected (Fig. 3.1).

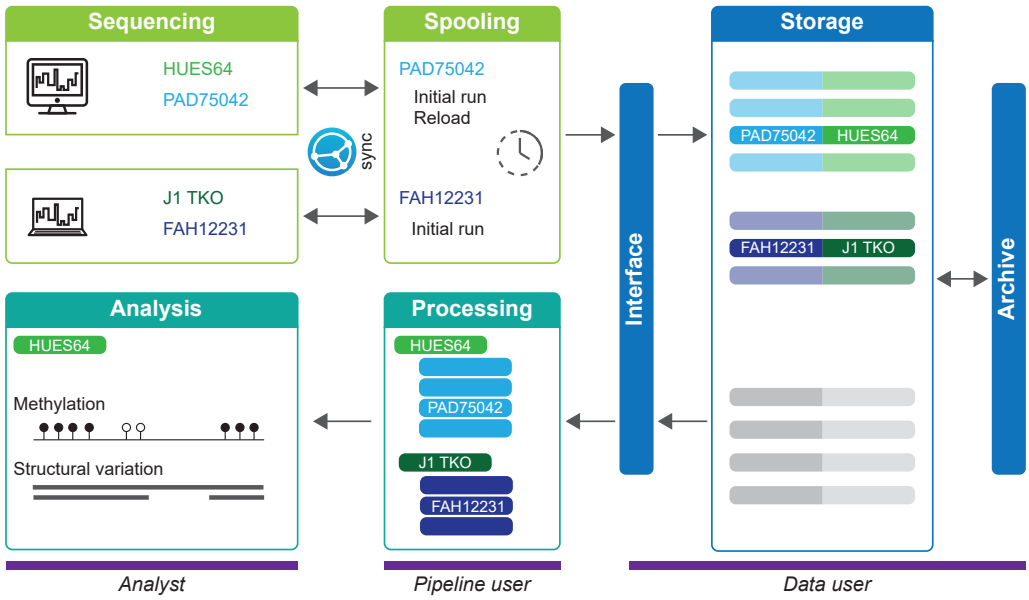


Fig. 3.1.: Data flow for multi-sequencer and multi-user setup: Raw data generated on local or remote devices is synchronized into a local spooling directory. Complete runs with meta information (PAD75042: flow cell ID, HUES64: sample name) are stored with read-only access and archived on tape. Processing combines multiple sequencing runs of the same sample and provides basic readout for downstream analysis.

The active storage contains one folder per sequencing run following a uniform naming pattern of date, flow cell ID, flow cell type, sequencing kit and a sequence of arbitrary user tags e.g.:

[3]https://github.com/nanoporetech/vbz_compression
[4]https://github.com/syncthing/syncthing

From the active storage, sequencing runs can be archived on e.g. tape drives for long term data retention. For the processing, sequencing runs are mapped into an interface (a directory with softlinks into the active storage). This additional layer allows the distribution of raw data across multiple physical devices, while maintaining consistent access. The raw data archive forms the basis for any downstream analyses and enables smooth re-processing of legacy datasets as soon as for example improved basecalling algorithms become available.

3.2.2 Encapsulation

The installation of experimental software packages still being actively developed with complex library dependencies can be time-consuming, but remains essential to make use of the current-generation nanopore analysis workflows. On a base level, *Nanopype* uses *Snakemake* rules to wrap the build from source and installation process of its dependencies and therefore does not require root privileges for the setup on common Linux and MacOS systems. Whenever available, wrapped software is built from versioned releases creating a frozen set with the respective pipeline release.

The internal wrappers are used to automatically build and deploy Singularity images for preset modules and pipeline versions. The images are automatically pulled once a module is used. This mechanism enables the complete function set of *Snakemake* and *Nanopype* while only requiring a system-wide Python and Singularity installation. In contrast to Docker, the execution of Singularity containers does not require root privileges on the target system at run time. An all-in-one Singularity container is provided, wrapping the entire pipeline into a single environment. Primarily aimed for stand-alone usage, Windows systems, and for initial testing, this method does not offer support for cluster computation.

While greatly simplifying the installation process, containerized software comes at the cost of the container size. Whereas the *minimap2* binary alone is 1 MB small, the size of the compressed *Nanopype* alignment image is 191 MB (v1.0.1) due to integration of alignment tools, samtools, libraries and the Ubuntu kernel. Especially within distributed environments where each cluster job needs to fetch these images from a file server, this may lead to noticeable overhead depending on the jobs core function.

3.2.3 Transparency

Common file formats in bioinformatics such as FASTQ or BigWig do not support the direct documentation of the applied workflow. An exception is the BAM alignment format, which can preserve the executed command in the @PG header line. Nonetheless, the source of the sequences remains obscure. Extensive and continuously updated documentation is required, especially when multiple parallel projects and workflows are involved.

Nanopype follows the Snakemake concept of output file driven computation: a single user command typically provokes transparent processing of any required intermediate result. Preexisting tools are integrated into consistent workflows and provide standard output formats to connect to workflows of established next-generation sequencing data analysis tools. The processing and subsequent output is intuitively organized in modules and underlying application directories (Fig. 3.2).

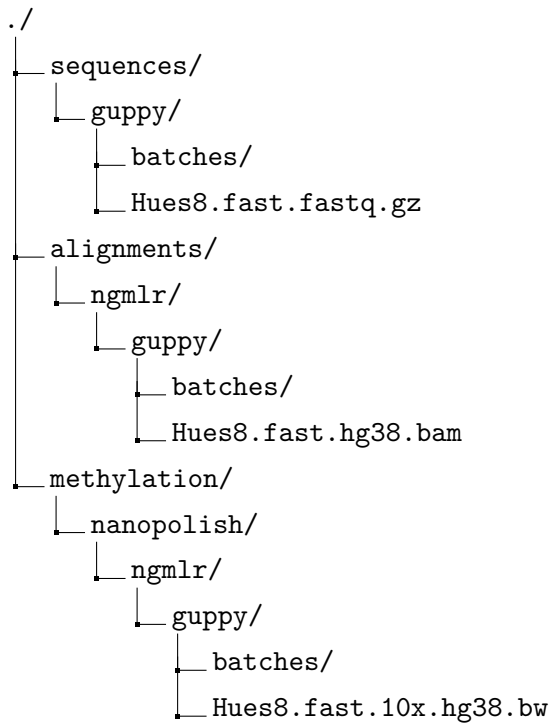


Fig. 3.2.: Typical *Nanopype* output directory structure after basecalling, alignment and methylation detection. Top level directories represent pipeline modules with the applied tools in subjacent levels. **Hues8.fast** is the user defined tag of the output, in this case for the HUES8 cell line and guppy basecalling in fast mode. **10x** is the coverage wildcard, indicating to filter for CpGs with at least 10 reads coverage. **Hg38** defines the reference genome and **.bw** the output file format.

The full reach of this concept becomes visible in the case of alternative workflows, re-processing and downstream analysis. Requesting for instance the output file

`methylation/nanopolish/minimap2/guppy/Hues8.fast.10x.hg38.bw` would detect the presence of basecalled reads for the tag **Hues8.fast** and run only the subsequent alignment and methylation detection. While the processing of `sv/sniffles/ngmlr/guppy/Hues8.fast.hg38.vcf.gz` would utilize the already existing alignments, the same workflow but for reference **hg19** would automatically generate the missing intermediate alignments. Lastly, any analysis using *Nanopype* as a pre-processing step can rely on filename patterns and the fixed output directory structure.

3.3 Modules

Nanopype's backbone consists of modules that resolve a specific task, like basecalling, alignment or further downstream analyses. If available, alternative applications are provided for the same task and grouped into a module with a coherent output format. Integrating first and foremost low-level nanopore data processing applications provided by ONT, established community developed software packages have been included in the first Nanopype release as well.

3.3.1 Basecalling

The basecalling module translates raw nanopore signals into nucleotide sequences and is utilized by most subsequent pipeline layers. With the initial release, we include the established packages *Guppy*, *Albacore*, and *Flappie*, all provided by ONT [37]. The default basecaller package is set to the recently released *Guppy*. *Albacore* is supported for backward compatibility but deprecated by ONT. The experimental *Flappie* is ONT's first DNA methylation-aware basecaller. It extends the usual four-letter nucleotide alphabet by a fifth letter for methylated cytosine in CpG contexts. For all basecallers, the output is the standardized FASTQ and supplemented by us with a basic quality control summary.

3.3.2 Alignment

The core functionality of the pipeline is the alignment of reads against a reference genome or draft assembly. Here, we provide three different aligners with distinct advantages, which make them favorable for different applications downstream. While *Minimap2* [41] is a fast, low memory footprint solution suitable for both DNA and RNA alignments, *GraphMap* [40] is a sensitive aligner but with comparably

high memory requirements. *NGMLR* [43] is the recommended tool for the structural variation module. Any combination of basecalling, alignment, and reference genome is supported and reports BAM format files.

3.3.3 DNA methylation

Sequencing without prior DNA amplification enables the direct readout of DNA base modifications. The current state of the art approach, *Nanopolish* [102] and the more experimental flip-flop basecaller *Flappie*, are incorporated into *Nanopype*. Subsequently, *Nanopype* splits *Flappie's* atypical sequence output into standard FASTQ and methylation status. DNA methylation at CpG dinucleotides of both tools is reported in a table format for single reads. Furthermore, we provide standard bedGraph and BigWig files for genome-wide methylation tracks and thus enable downstream processing and visualization using established workflows, e.g., calling of differential methylated regions and comparison to bisulfite-sequencing.

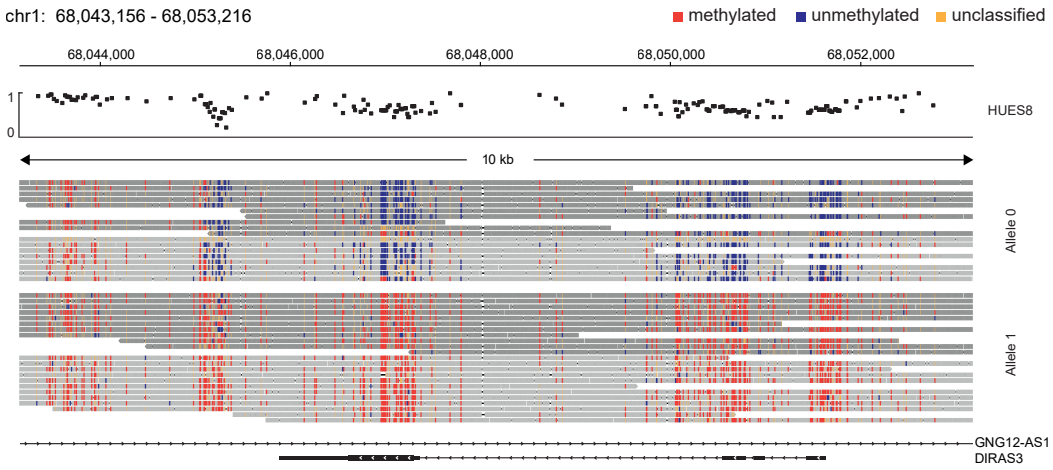


Fig. 3.3.: *Nanopype* single-read methylation track of nanopore sequenced HUES8. Mean methylation track of CpGs with $>=10x$ coverage. Single-read methylation computed from original alignment by substituting unmethylated Cs to Ts (imitated bisulfite conversion, blue). Unclassified sites (abs. *nanopolish* log-likelihood-ratio < 2.0) are substituted with As and appear as mismatches (orange). All remaining mismatches in the nanopore reads are replaced with the genomic sequence. Shown are reads overlapping DIRAS3, an imprinted gene on genome build hg38, grouped by allele and shaded by strand.

3.3.4 Structural variation

Detection and characterization of structural variation play a central role in cancer research and population genetics. Long read sequencing particularly facilitates investigation of variants with unprecedented accuracy and resolution. Therefore,

Nanopype encompasses the variant callers *Sniffles* [43] and *SVIM* [64] and provides output in the standard variant calling format (VCF).

3.3.5 Transcriptome

Another application of the long read nanopore technology is sequencing of cDNA and RNA molecules directly. For instance, recovery of full-length transcripts enables, the detection of alternatively spliced isoforms and is implemented in Nanopype using the *Pinfish* package. The output of polished transcripts is provided in the GFF format.

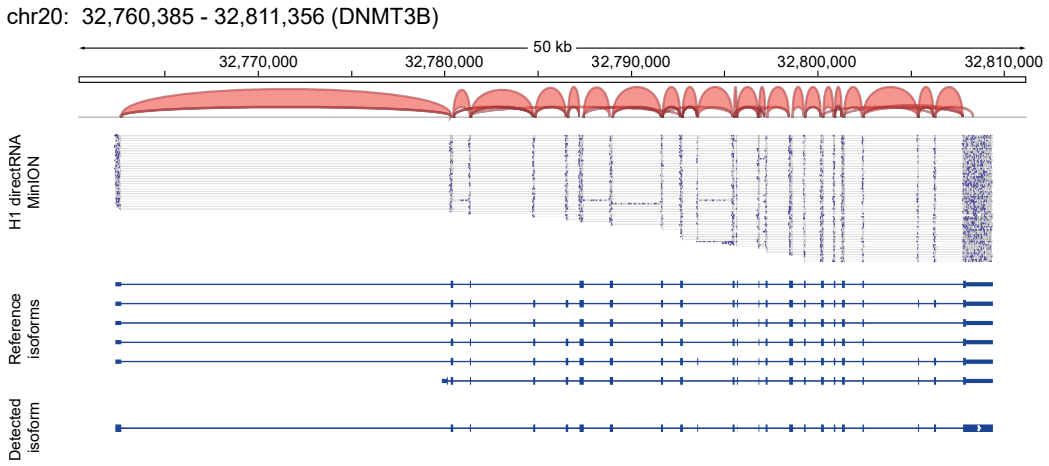


Fig. 3.4.: Nanopore direct RNA sequencing of H1 from one MinION flow cell. Shown are *minimap2* splice-aligned reads spanning DNMT3B, obtained by using the RNA sequencing without amplification. The isoform detected from the *Pinfish* package matches one of the reference isoforms.

3.3.6 Genome Assembly

De novo assembly of genomes requires only the long read sequences and benefits barely from the integration into a pipeline. Due to its own cluster back end, the still widely used *Canu* [24] has not been included. The recently published *Flye* [59] and *wtdbg2* [60] assemblers provide comparable continuity of the output genome while reducing the required compute time. Both are part of Nanopype, enabling basic assemblies already from 20-30x coverage.

Complemented by *SAMtools* [114], *BEDtools* [115] and *UCSCtools* [116] our pipeline establishes a comprehensive framework for ONT sequencing data processing.

3.4 Installation

The installation process of dependencies is a crucial, though sometimes neglected step to deploy reproducible workflows. A commonly found bypass is to require tools to be discovered through the user's PATH variable, with a remaining uncertainty of particular versions and thus limiting the outcome to the order but not necessarily exact behavior of applied tools. We address this issue by providing installation wrappers, enforcing tool versions to be bound to each individual pipeline version. Programs and files are identified by their absolute path, making the approach robust against changes of environment variables between local and cluster servers. Within *Nanopype*, consistent versioning of source and container builds ensures reproducibility independent of the installation method.

3.4.1 Source

We utilize the workflow engine Snakemake to automate cloning and building of open source software. The build from source is the most customizable setup option and moreover allows the integration into complex, preexisting environments. The snakefile *rules/install.smk* in the *Nanopype* repository contains installation wrappers for all included tools, except proprietary software like the basecaller *albacore*. For example, the following command would clone and build the *minimap2* binary required by the currently installed pipeline version. A more detailed description of the Snakemake command line is given in section 3.5.

```
snakemake --snakefile ~/src/nanopype/rules/install.smk --
    directory ~/ minimap2
```

Listing 3.1: Snakemake tool installation example

After completion, the home directory will contain the common *src*, *lib* and *bin* folders and the executable *~/bin/minimap2*. As an example, the source of the *minimap2* build rule is given in listing A.1.

A special case when building from source is formed by heterogeneous server cluster environments in combination with vector instructions such as MMX, SSE and AVX. Depending on the software, the support of these parallel instructions is determined at build- and not at run time. Compiling the toolset of *Nanopype* on a modern CPU with e.g. AVX2 can cause program crashes (Signal 4, illegal instruction) on other

older nodes. A simple solution can be building the pipeline on a node with the largest common subset of advanced CPU instructions.

3.4.2 Container

Docker and Singularity enable the encapsulation of tools and dependencies into virtualized software containers. Packaging libraries and data, these containers can be executed on any operating system with just the virtualization set up. While the source installation method of *Nanopype* aims to minimize dependencies, a functional C/C++ tool chain and a basic set of development libraries are required to build all parts of the pipeline. To further reduce both, target site dependencies and installation time, we provide pre-build Singularity containers per pipeline module and as all-in-one.

Nanopype module images are build and tested on Travis-CI and deployed as Docker containers to *https://hub.docker.com/u/nanopype/*. Being executed by *Singularity* in the *Snakemake* backend, both the Docker and Singularity container formats are supported, allowing a choice based on personal preferences. Webhooks from Github trigger builds on each push to master and development branch. Tagged commits (e.g. *v1.0.1*) to the master branch are build as tagged images. Upon execution the pipeline version is inferred from the git tag and associated images are pulled as needed for the workflow.

In order to minimize the size of each container, all *Nanopype* modules rely on a staged build process. A *base* and a *build* container are set up in a first step, one with the pipeline and its basic dependencies, the other with build tools and libraries. Within each module's Dockerfile two separate stages are configured. The first build stage inherits from the *build* container. Bioinformatic tools required by *Nanopype* are compiled and linked in this stage. As an example, the build process for the alignment module is shown in listing 3.2.

```
# BUILD STAGE
ARG TAG=latest
FROM nanopype/build_bionic:$TAG as build_stage

## run setup rules
RUN mkdir /build
WORKDIR /app
```

```
8  RUN snakemake --snakefile rules/install.smk --directory /
       build alignment
9
10 # PACKAGE STAGE
11 FROM nanopype/base_bionic:$TAG
12 MAINTAINER Pay Giesselmann <giesselmann@molgen.mpg.de>
13
14 ## copy packages from build stage
15 COPY --from=build_stage /build/bin/* /usr/bin/
16 WORKDIR /app
17 # default entrypoint is /bin/sh
```

Listing 3.2: Staged Docker build

A subsequent packaging stage inherits from the *base* container and copies all needed binaries from the build stage and requires only the installation of runtime libraries for dynamically linked executables. The build stage is dropped in the final compressed image.

The all-in-one container works in a similar way, by first pulling all module containers into intermediate layers, copying their binaries and squashing everything into a single image. The cascaded build process from source over module to all-in-one container reduces redundant code to a minimum and enhances the maintainability of the pipeline. Thus, to update or add additional tools, only the source build rules need to be edited.

3.4.3 Configuration

The last part of the installation is the site dependent configuration. Nanopype has two configuration layers: The central environment configuration *env.yaml* covers application paths and reference genomes and is set up independent of installation method and operating system once. The environment configuration is stored in the installation directory. For each project an additional workflow configuration is required, providing data sources, tool flags and parameters. The workflow config file *nanopype.yaml* is expected to be found in the root of each processing directory. Configuration files are in .yaml format; examples with default values can be found in the pipeline repository.

If a compute cluster is available, the respective Snakemake configuration is only needed once per Nanopype installation. Already available presets support a custom scheduler called mxq and the common SLURM scheduler. Presets are stored in the profiles folder of the repository and can be extended by pull requests through the community. A compute rule can, depending on the number of threads assigned, define memory and time requirements. These parameters can be forwarded to the cluster scheduler upon execution. Both memory and run time are conservatively pre-configured but can be adjusted via individual offset and scaling parameters in the environment configuration.

Nanopype makes extensive use of the Snakemake shadow directory mechanism. Per default, rules are executed directly in the working directory. By specifying a shadow-directory, the input of supporting rules is linked into a temporary location and only expected outputs are copied to the working directory. Intended for the isolation of tools producing temporary intermediate outputs, this concept can also be used to move I/O heavy computations to local hard drives on each compute node and thus reducing the load on network file systems. Given that each compute node in the cluster has a local disk mounted as e.g. */scratch/local*, the Snakemake shadow-prefix would be set to this path.

3.5 Usage

Due to the functional range of *Nanopype*, dependent on the operating system and selected installation method the setup can require advanced system administrative knowledge. However, after deployment, the subsequent usage is straightforward, given basic command line understanding. Complete *Nanopype* workflows can be executed with a single concise command line call. For instance, local processing of multiple flow cells into a collective genome-wide methylation track of at least 5x coverage on reference hg38 requires only the following call:

```
1 snakemake --snakefile ~/nanopype/Snakefile methylation/
    nanopolish/ngmlr/guppy/Hues8.5x.hg38.bw
```

Listing 3.3: Snakemake example

This command invokes basecalling, alignment and methylation detection using declared tools without further user interaction. The basecalling and alignment outputs are kept and can be reused to avoid redundant processing.

The output path encodes the applied toolchain, the filename is composed of the tag and a sequence of workflow dependent wildcards. For the given example, **Hues8** is the name of the cell line, while **5x** and **hg38** are parsed by the pipeline to determine minimum coverage and reference genome. Different tags can be used to test different settings on the same data set e.g. basecalling with guppy in fast and high-accuracy mode could be indicated by **Hues8.fast** and **Hues8.hac**.

Tags are global, changing only the alignment settings and re-running a workflow would still trigger a new basecalling, since sequences for the new tag are not available.

3.5.1 Batch processing

Nanopype exploits the storage of raw nanopore data into batches of single reads. The automatic distribution of workflows into independent compute portions enables efficient handling of high-throughput experiments. This feature becomes particularly relevant for scaling in cluster environments and, most importantly in case of terminated or failed jobs. As a result, only failed batches require reprocessing by resuming the workflow from where it left off, using the same command, which enhances the overall error robustness. *Nanopype* keeps the dataset separated as long as possible: For instance does the methylation detection of a single read only require its sequence and alignment, while the structural variant detection requires all alignments to be merged into a single sorted file.

When working with batches of nanopore reads, two limitations of the underlying operating system require special handling: The maximum command line length and the limit of parallel opened files. Taking the example of sequence alignment and sorting, a naive merge of all batches into a single BAM file would call *samtools merge* followed by the list of independently computed alignment batches. For high throughput experiments, this is likely to break either the maximum command argument length (getconf ARG_MAX, on ubuntu e.g. 2097152) or the maximum allowed open files per process (ulimit -Sn, on ubuntu e.g. 1024). The former is solved by reading from files of filenames (.fofn) and piping data through stdin, the latter by first merging multiple times with the maximum opened files and a final merge of these intermediate results.

3.5.2 Barcoding

Barcoded sequencing allows pooling of multiple samples on a single flow-cell. Thus, sequencing of comparable small bacterial genomes can be efficiently parallelized to use the available sequencing depth optimally. The corresponding demultiplexing is a special transparent module in *Nanopype*. Using *Deepbinner* [117] on signal or *Guppy* on sequence level, it assigns a barcode label to each individual read. The following example illustrates the command to only process reads of barcode NB01 from the ONT native barcoding kit:

```
snakemake --snakefile ~/nanopype/Snakefile sequences/
   guppy/Ecoli.NB01.fastq.gz
```

Listing 3.4: Nanopype demultiplexing

The pipeline automatically scans the previously introduced tag for substrings indicating the usage of a barcode. Barcodes are specified through an additional config file *barcodes.yaml*. Indexing the content of both packaged and multi-FAST5 output enables the fast retrieval of individual reads by their ID. The demultiplexing module first generates batches of read IDs per barcode and then temporarily extracts those reads for downstream processing.

3.5.3 Logging and Reports

Finally, an extensive logging of both, job specific output on each compute node and a summary of configuration values, ensures the complete documentation of the workflow. The quality control of the sequencing is enabled by automatic pdf reports containing basic statistics as read lengths, mapping rates and coverage. The report is obtained by the following command and contains dynamic sections for each of the pipeline modules.

```
snakemake --snakefile ~/nanopype/Snakefile report.pdf
```

Listing 3.5: Nanopype report

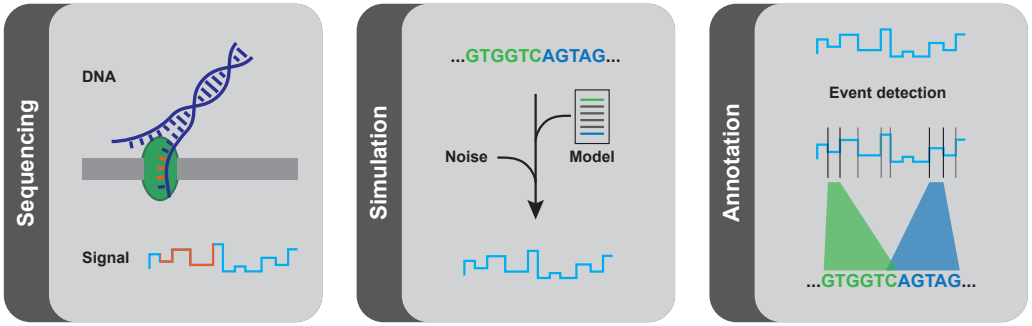
An example report is attached in supplement A.1.

3.6 Summary

Nanopype is a modular and easy-to-use data processing pipeline with a detailed online documentation, specifically designed to handle nanopore sequenced long read data. *Nanopype* provides end-to-end processing of the raw sequencer signal into standard data formats and consequently closes the gap to downstream next-generation sequencing algorithms. Single command invocations of entire workflows reduce the hands-on-time for users to receive the desired output. Implicitly, this also lowers the potential of user mistakes and deviations in processing of multiple data sets. Consequently, workflows are easier to reproduce with fixed versions among datasets or repeated with improved tool releases on existing ones. *Nanopype* is implemented as a Python package and additionally provides pre-built and versioned Singularity and Docker images, making it favorable for effective usage in cluster and single computer environments. The pipeline design ensures portability and version controlled usage of the implemented tools, to enable consistent results across platforms and laboratories.

4 Nanopore Signal Analysis

The primary output of third-generation nanopore sequencing devices is the raw ionic current measured as a proxy signal for a DNA or RNA strand passing through the pore. An initial processing step of particular importance is the basecalling, translating the raw signal into the respective genomic sequence. Common workflows such as genome assembly, structural variant or isoform detection typically rely on sequence inputs only. However, the error rate and run time of state of the art basecalling algorithms motivate ongoing research on raw signal processing itself. Furthermore, the sequencing without amplification preserves signatures of modified bases in the signal of DNA and RNA samples. Prominent applications based on raw nanopore signals include methylation detection, barcode demultiplexing and real-time alignment for selective sequencing. Challenges in the raw signal analysis arise from noise induced by measuring currents in pico ampere ranges and time-warping, the uncertainty of how long the molecule resides stationary in the pore before being advanced by the motor protein.



The following chapter serves as a transition from basic data handling and processing using *Nanopype* in chapter 3 and introduces the underlying algorithms for the signal driven repeat detection *STRique* in chapter 5. After a brief **background** in section 4.1, this chapter covers the raw signal **simulation** from known sequences in section 4.2. Noise and time warping are addressed by the **normalization** followed by **event detection** and **annotation** of raw nanopore reads with reference sequences in sections 4.3 and 4.4.

4.1 Background

In contrast to second generation sequencing, the main output of nanopore sequencing is the ionic current measured on the device while a molecule is passing through any of its pores. The characteristics of this signal are determined by pore and chemistry version, currently R9.4.1, as released by ONT in October 2016. In March 2020 ONT released the R10.3 (dual reader head), promising better consensus accuracy, particularly for homopolymer stretches. In terms of throughput per flow-cell, the R10 is slightly lacking behind R9 [118]. Independent of the pore version, the storage of the raw signal is valuable, as constant development by ONT and researchers in the community provides enhanced readouts from existing sequencing data.

From a technical perspective, the ionic current is sampled at 4kHz and saved as a 16-Bit unsigned integer array to FAST5 files. In combination with the sequencing speed of ~450nt/s, determined by the chemistry, this results in a square wave-like signal with a median of 9 measurements per nucleotide. Being a specification of the HDF5 file format, FAST5 files can be inspected using GUI and command line tools like *HDFView* and *h5dump*. File access from custom software is available for C/C++ programs by linking against the HDF5 library[1] and from python using the *h5py* package. Additionally, to work with the most recent data sets, the VBZ compression plugin[2] is required.

While the raw nanopore signal is utilized in some bioinformatic applications [102, 117, 119], the set of methods providing frameworks is currently limited to *Tombo*[3], the successor of *Nanoraw* [120] and the recently published *SquiggleKit* [121]. The motivation to develop and use custom code for the raw signal analysis is driven by the lack of sufficiently effective and customizable algorithms. For example, *Tombo* is hard-coded to usage of *minimap2* alignments and writes its output in form of event tables into the FAST5 files. While technically supported by the FAST5 format, writing of analysis results into the sequencing data is generally undesirable in a multi-user environment. It causes, in the case of FAST5 files, excessive disk usage since new, but also overwritten data sets are appended to the end of the file. *SquiggleKit* so far only provides basic data indexing and signal alignment, mostly to identify barcodes in raw reads.

[1] https://bitbucket.hdfgroup.org/projects/HDFFV/repos/hdf5/browse
[2] https://github.com/nanoporetech/vbz_compression
[3] https://github.com/nanoporetech/tombo

Publicly available, highly customizable methods in the form of a generic API are still lacking, demanding the development of an in-house framework to handle raw nanopore reads. Future work aims to include advanced normalization methods like iterative re-normalization after signal alignment as described in [34] and hardware acceleration, comparable to the GPU implementation of a banded signal alignment in the *f5c* package [122].

4.2 Simulation

Third generation sequencing data can be simulated on different levels by generating sequences with realistic read length and error rate distributions or raw signal traces, ideally indistinguishable from measured ones. Simulated reads of both types are helpful to develop or test applications under controlled conditions. Two state of the art simulators are *simulatION* [123] and *deepSimulator* [124], following different concepts to generate FAST5 files of simulated reads, which can be processed with common workflows. While the former utilizes exiting models of level and noise, the latter uses a LSTM neural network to generate the signal from reference sequences. Nonetheless, generation of realistic signal traces in memory can already be achieved with a pore model and an event length distribution.

A pore model describes the mapping from sequence to expected signal. The ionic current level in the nanopore is determined by the local sequence context, termed kmer. Current models utilized by e.g. *Nanopolish* are built of the mean and standard deviation of signal levels per 6mer. The distribution of all levels, in this case 4096, is shown in Fig. 4.1 a, emphasizing the multi-modal characteristic of the R9.4 pore.

In general, the nanopore is sensitive for several base modifications, which are preserved when sequencing without prior amplification. A secondary pore model can be derived from sequencing e.g. 5-methylcytosine modified DNA, enabling the discrimination of native and 5mC on signal level. The model difference between native and 5mC for single-CpG containing 6mers is illustrated in Fig. 4.1 b, showing a pronounced difference on the last position (Native and methylated model taken from *Nanopolish*). The combined differences of all CpG overlapping kmers allow the discrimination between methylated and native DNA.

In addition to the signal level, the event length or dwell time of the molecule in the pore needs to be modeled. It is not fully understood, to which extend the behavior of the motor protein controlling the sequencing speed is a random or

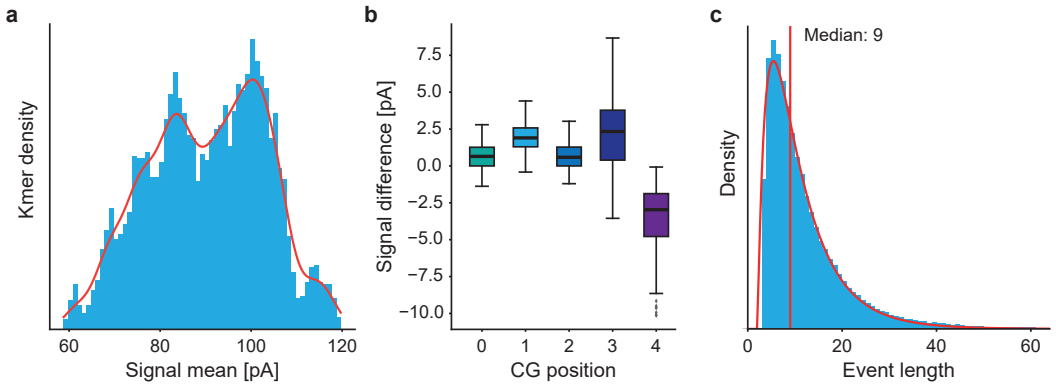


Fig. 4.1.: Pore model and event lengths: **a**, Density plot and kernel density estimation of expected signal levels per kmer (k=6, n=4096, mean ionic current in pA). **b**, Expected signal difference of native and 5mC modified DNA for 6mers with single CpG (n=256 per position) depending on the CpG location. **c**, Event length (samples per kmer) density plot from signal alignment (red: approximation by generalized gamma distribution with a=4.3, c=0.6, shift=2, scale=0.7).

sequence context dependent process. Aimed solely towards the development of signal alignment methods, event lengths are randomly sampled from a generalized gamma distribution in this work. The distribution is derived from in-house sequencing data using the event detection described in section 4.4 and shown in Fig. 4.1 c. The measured event median of 9 samples per 6mer matches the expected value for a sequencing speed of 450nt/s sampled at 4kHz.

With pore model and event length distribution, simulated raw nanopore signals can be generated from any given target sequence. The mean event level, a simulated and corresponding stretch from a real nanopore read are depicted in Fig. 4.2.

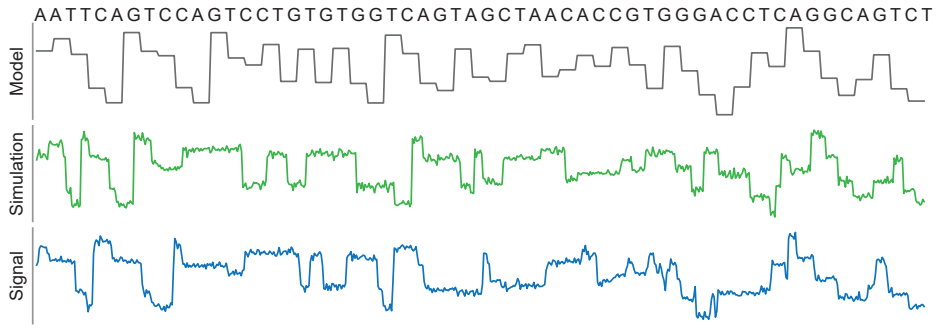


Fig. 4.2.: Basic nanopore signal simulation from pore model and event lengths: Mean ionic current levels per kmer for short sequence (top), simulated nanopore signal with noise and random time warping (center) and raw signal fragment extracted from real nanopore read covering the same sequence (bottom).

Both, the published and the proposed simplified simulation do not mirror every detail of actual sequencing data and can therefore only serve as an approximation. Examples of missing technical artifacts are rare spikes of single values and tem-

porarily stalled reads with hundreds of samples from the same kmer. Lastly, motor protein and analog-digital converter are not synchronized, frequently resulting in measurements on the rising or falling edges between events.

4.3 Normalization

A robust normalization is a substantial initial step during nanopore signal processing, impacting any downstream method and readout. Baseline for the following normalization is the raw ionic current measurement, saved by the sequencing software *MinKNOW* as an unprocessed unsigned integer with a typical numeric range from 300 to 700 (cf. Fig. 4.4 top track). Before any downstream processing, we apply a median filter with a sliding window of length three to remove signal spikes and reduce the overall noise. Factors further affecting the signal levels are: Individual offset and scale induced by the electronic circuit per pore and an uneven sequence composition depending on the genomic context resulting in biased sampling from the multi modal pore model.

Offset and scale for normal distributed signals (X) can be compensated by using a z-score normalization:

$$X_{norm,z-score} = \frac{X_i - \mu}{\sigma} \tag{4.1}$$

Suggested by *Nanoraw* and claimed [120] to be more robust against the underlying distribution is a normalization with median offset and median absolute deviation (MAD) scale:

$$X_{norm,MAD} = \frac{X_i - median(X)}{median(|X_i - median(X)|)} \tag{4.2}$$

With regard to the signal driven repeat quantification in chapter 5, we propose a novel strategy combining a min-max and quantile normalization. Subtraction of the signals 0.025 quantile (Q) and scaling by its interquantile range is expected to be more stable, especially for biased signal distributions:

$$X_{norm} = \frac{2 \cdot (X_i - Q_{0.025})}{Q_{0.975} - Q_{0.025}} - 1 \tag{4.3}$$

The suitability of normalization algorithms applied to real nanopore data can, due to the absence of true scale and offset values, only be implicitly rated by the performance of subsequent methods. Therefore we evaluate the proposed min-max normalization against the other methods on simulated signals and assess their ability to minimize the distance between simulated and normalized signals. Taking the different statistical characteristics of each normalization output into account, we first generated specific pore-models for each method, mapping the signal levels from pico ampere to e.g. the [-1:1] interval of the min-max method. Next we sampled random fragments from the human genome hg38 with lengths from 1kb to 10kb. Nanopore signals were simulated with event levels from each pore model and event lengths only sampled once per read to be comparable across all three methods.

The sequence complexity is expected to influence the normalization due to a biased sampling from the multi-modal pore model. The complexity is measured as observed kmers per read, divided by the total kmer count and its distribution is illustrated in Fig. 4.3 a. While generally rising with increased read length, none of the simulated reads contains all possible 4096 6mers.

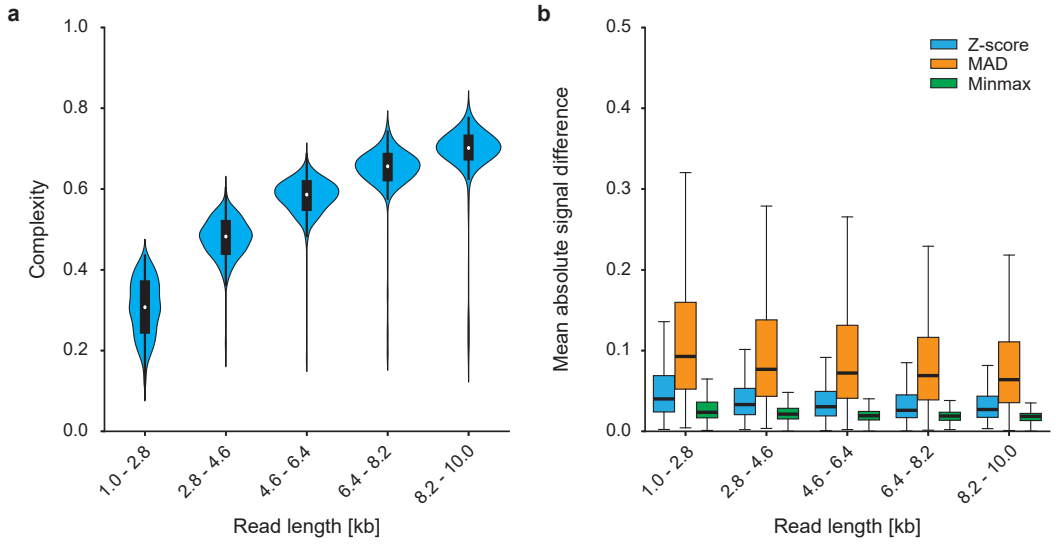


Fig. 4.3.: Comparison of different normalization strategies on simulated raw nanopore signals. **a**, Sequence complexity as fraction of unique kmers per read divided by total kmer count (k=6, n = 487, 536, 485, 485, 507 reads) Data in a are presented as violin plots with overlayed boxplots. **b**, Mean absolute difference between simulated and normalized signal grouped by read length and normalization method. Read lengths and counts as in a, with matching event lengths used across all normalization methods. Data in a-b as boxplots (centerline, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range).

Ideally, a normalization method applied to a signal from a normalized pore-model is expected to be transparent. Nonetheless, the observed mean absolute differences of simulated and normalized signals show notable deviation, especially and surprisingly in the case of the MAD method (Fig. 4.3 b). With overall smallest differences and es-

pecially robust against low complexity regions, indicated by less pronounced outliers, the proposed min-max normalization appears to outperform previous methods. We conclude, that the min-max normalized signal with most values in the range [-1:1] maintains the characteristics of the nanopore, is comparable to simulated reads using a pore model and is therefore used in this form for the repeat quantification in chapter 5.

For the purpose of signal alignment and event segmentation, we propose two additional steps of histogram equalization and morphological smoothing [125]. Commonly used to enhance the contrast of digital images, a histogram equalization can be used to project the multi-modal distribution of mean event levels to a more uniform distribution (Fig. 4.1 a, Fig. 4.4 bottom track). The histogram equalization would be affected by uneven sequence compositions and is therefore learned on a random set of long reads, before being applied individually. Consequently the static multi-modal pore model can be normalized without sensitivity to read specific signal compositions. To further reconstruct the expected square wave like signal, we apply a morphological noise removal, specifically an opening followed by a closing with a structuring element of length three.

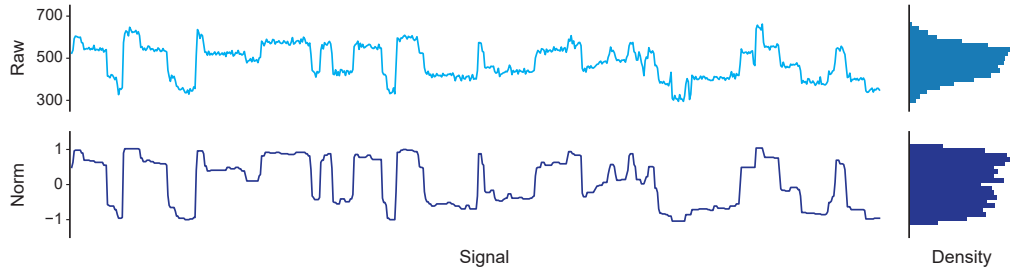


Fig. 4.4.: Signal normalization and histogram equalization followed by morphological noise reduction on raw nanopore signal traces. The figure shows signal segments over a 50nt window in combination with density plots of raw and normalized values over the entire read.

The final result of normalized and filtered signal, with a near uniform distribution of values across the full signal range, is the basis for the alignment and event detection described in the following section.

4.4 Signal Alignment

For applications, where neither signal nor sequence alone enable the intended readout, the sample wise annotation of raw signals with the respective reference sequence is required. Affected by the basecalling error of around 5-10% (cf. Fig. 2.9 b), the usage of the reference sequence after read alignment is preferable

over the read sequence for this purpose. The following two sections describe the **segmentation** and alignment of nanopore signals to extract any region of interest from a read, followed by the precise **annotation** of signals with reference sequences and base modifications using Hidden Markov Models (HMM). The former is the enabling step for the latter, as HMMs become computationally very expensive with increasing numbers of hidden states and sequence lengths.

4.4.1 Segmentation and Event Detection

A semi-global signal alignment is calculated between the normalized raw signal and a simulated signal with a constant event length of either the complete reference span or a query region of interest within the read. Primarily to extract larger regions of interest from a raw signal, we specify a template function of the *SeqAn2* [126] library to perform a distance-based semi-global signal alignment. Specifically, we utilize the *globalAlignment* function with a float32 data type, an affine gap penalty and the following score function:

$$s_{i,j} = max \begin{cases} c - |x_i - y_j| \\ 0 \end{cases} \tag{4.4}$$

The absolute signal difference per position is transformed into a score by subtracting it from a constant c. The score is capped at zero as a lower boundary, resulting in a fixed minimum mismatch score, independent of the actual signal difference. Gap costs within signal and simulation are configured differently. Both, gap open and gap extension penalty in the signal are set to a small negative value with begin and end gaps being free (semi-global alignment). Gap open and extension in the simulated reference signal are scored with a penalty an order of magnitude larger. The rationale behind is, that gaps in the signal are expected and introduced by simulated events of length one, being stretched to the observed lengths. Gaps in the simulated reference signal are expected to be very rare and result from sequence stretches not being observed in the raw signal.

The described approach is limited by the modeling of event lengths with affine gap costs and more severe by the quadratic time and memory complexity of the semi-global alignment with respect to the read length. Read lengths of multiple hundred kilobases make direct alignments in signal space increasingly ill-suited to annotate the full read. Implementations in *Tombo* and *Nanopolish* [102, 122] address

this issue with banded alignments, without appreciating the underlying problem of an oversampled square wave like signal.

Inspired by a general concept for the discretization of time series into symbolic strings [127], we further improve our signal alignment by an event compression and discretization step. Exploiting the distinct steps between events after histogram equalization and morphological noise reduction, we apply an image processing edge detection on the normalized signal using a one-dimensional convolution with the kernel [-3 0 3]. Intermediate segments are summarized into an event table with their lengths and mean signal levels (Tab. 4.1). The previously described uniform distribution of signal values enables a discretization into equally represented symbols, by splitting the signal range into evenly sized bins. A lookup table mapping reference kmers from the pore model to the same symbol space can be pre-computed. Lastly, any generic sequence alignment library can be used to align both symbolic sequences without handling the nanopore specific time-warping.

Tab. 4.1.: Event table of raw nanopore signal with reference sequence annotation.

	mean	event length	seq. offset	kmer
		...		
E_n	1.332	15	3205	AGTCCA
E_{n+1}	0.981	22	3206	GTCCAG
E_{n+2}	-1.058	6	3208	CCAGTC
E_{n+3}	-1.662	17	3209	CAGTCC
E_{n+4}	1.360	7	3210	AGTCCT
E_{n+5}	0.664	33	3211	GTCCTG
E_{n+6}	0.107	11	3212	TCCTGT
E_{n+7}	0.844	4	3213	CCTGTG
E_{n+8}	1.362	38	3213	CCTGTG
		...		

Here, the event alignment is implemented using the python module of the *Edlib* [128] package. Based on Myer's bit-vector algorithm and with alignment path traceback in linear memory (Hirschberg's algorithm), the library supports symbol sequence alignments over alphabets of up to 256 characters. The overall sensitivity of the method depends on the edge detection threshold, resulting in an over- or under-detection of events and the size of the discretization alphabet. In this work, we align event sequences represented by an alphabet of 12 characters. Furthermore, *Edlib* allows to extend the character equality matrix, which we utilize to treat symbols to be additionally equal with adjacent symbols in terms of the signal level they represent.

A small extract of the resulting event annotation is shown in Table 4.1. Highlighted are a row followed by a missed event (deletion, orange) and two rows belonging to the same reference kmer (insertion, green). Taken together, the event detection and alignment maps 90% of the events to a unique kmer, 5% are over segmented and the remaining 5% are distributed to the most critical errors of single and consecutive reference kmers without signal observations.

The above signal alignment enables reference guided extraction of any region of interest from long nanopore reads. Yet, for the analysis of base modifications encoded into the signal or regions diverging from the reference genome, a more sophisticated model is needed.

4.4.2 Annotation

Already proposed for the analysis of nanopore signals from previous pore generations, Hidden Markov Models are a powerful resource to map signal onto sequence features [129]. In contrast to the discrete emission distributions, used in for example sequence motive detection, the hidden states model in this case the sequence context depending signal observed in the pore, overlayed by Gaussian distributed noise.

ONT's first generation of basecalling algorithms used HMMs with hidden states derived from kmer signal levels and all possible transitions between consecutive overlapping kmers. The Viterbi path through such a model is the most likely state- and thus genomic-sequence given the observed signal. Whereas outperformed by recurrent neural networks for basecalling, HMMs still provide state of the art performance for base modification detection such as in *Nanopolish*. For the repeat analysis in chapter 5 we use profile Hidden Markov Models closely following the concept of modular blocks described in [129].

Starting with a target sequence, a series of match states from overlapping kmers and with normal distributed signal emissions forms model's expected path. Additional insertion states with uniform emission distributions over the whole signal range compensate observations outside of the match state distributions. Lastly, silent deletion states allow the model to skip parts of the profile sequence without corresponding observations in the signal (Fig. 4.5 a, b).

Simply chained by single transitions to the four outer silent states (s1/2 and e1/2), the modular profile HMM blocks can be used to build more complex architectures as illustrated in Fig. 4.5 c. The first example shows a possible base modification detec-

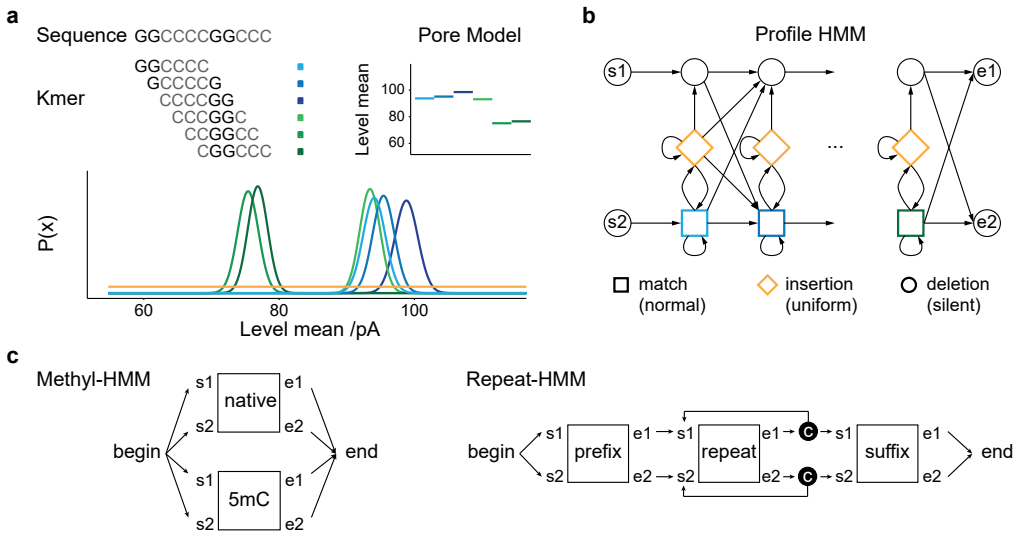


Fig. 4.5.: Nanopore signal alignment with HMMs: **a**, Consecutive overlapping kmers from profile sequence with corresponding pore model levels and expected emission distributions in the pore. **b**, Nanopore signal profile HMM with normal distributed match states and uniform distributed insertion states. **c**, A methylation detection HMM with emission distributions for native and 5mC modified DNA. A compound profile HMM of prefix, a single repeat and the suffix sequence with dummy states counting transitions through the repeat module.

tion architecture with two profile HMM components derived from the same sequence, but with emission distributions centered around expected signal levels of native and, for instance, 5-methylcytosine modified DNA. Computing the log-probability and Viterbi path through this model yields the most likely hidden state sequence of one branch, given an observed nanopore signal. The second example addresses the case of heterogeneous genomic regions, where reference and sequenced individual diverge. In-depth introduced in the following chapter, short tandem repeats are sequence fragments with varying repetition counts per individual. The illustrated compound profile-HMM is anchored by stable and known genomic prefix and suffix sequences, framing a single instance of the repeat. Feedback transitions around the repeat enable the model to adjust to any repeat length. In this case, the Viterbi path as the most likely state sequence given the observed signal can be used to quantify the length of the tandem repeat.

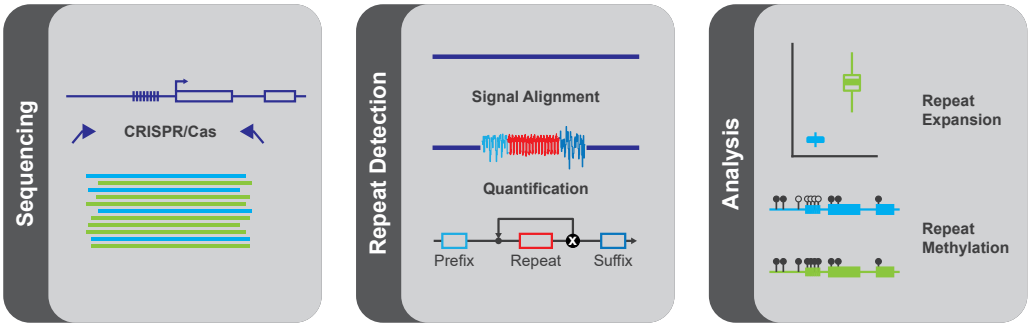
4.5 Summary

Taken together, this chapter introduces signal processing methods required to integrate the genomic sequence with additional information encoded in the raw nanopore signal. Normalization methods reversing scale and offset induced by the sequencer have been compared and form the backbone for subsequent analysis

steps. The advancement of existing banded signal alignment algorithms by event compressed symbolic sequences make this framework fast and memory efficient. Lastly profile Hidden Markov Models are introduced as an universal but also computationally most expensive method. In the following chapter, the above methods are applied to analyze the biological phenomena of expanded short tandem repeats in clinically relevant disease contexts.

5 STRique Repeat Detection

Expansions of short tandem repeats are genetic variants that have been implicated in several neuropsychiatric and other disorders, but their assessment remains challenging with current polymerase-based methods. Here we combine a CRISPR-Cas-based enrichment strategy for nanopore sequencing with an algorithm for raw signal analysis. Our method, termed *STRique* for short tandem repeat identification, quantification and evaluation, integrates conventional sequence mapping of nanopore reads with raw signal alignment for the localization of repeat boundaries and a Hidden Markov Model based repeat counting mechanism. We demonstrate the precise quantification of repeat numbers in conjunction with the determination of CpG methylation states in the repeat expansion and in adjacent regions at the single-molecule level without amplification. Our method enables the study of previously inaccessible genomic regions and their epigenetic marks. *STRique* is available at *https://github.com/giesselmann/STRique*.



Note: This chapter is based on the publication P. Giesselmann et al. *Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing*, Nature Biotechnology, 2019 and contains text and figures from the original paper.

The chapter starts with a brief **background** in 5.1, followed by the evaluation of **sequenced based** repeat analysis in 5.2.1. The development of an accurate **signal based** method is described in 5.2.2 and applied to patient samples with **c9orf72** and **FMR1** repeat expansions in 5.2.3. Finally the DNA methylation detection on repeat and surrounding sequence is shown in 5.3.

5.1 Background

Short tandem repeats (STRs), also called microsatellite repeats, form the shortest class of tandem repeat elements in the genome. They consist of short nucleotide sequences (2 to 6 bp) concatenated without other sequence fragments in between. Distinguished by length, the next larger tandem repeat class, termed variable nucleotide tandem repeat (VNTR) or minisatellite repeat follows the same repeat pattern but from larger fragments (>6 bp). VNTRs are for instance used as genetic fingerprints in forensic crime investigations due to their highly variable lengths within populations [130]. Lastly alpha satellite repeats are a class of large ~171 bp repeat elements structuring the centromeric regions of the human genome [131].

The expansion of unstable genomic STRs is of particular interest as it causes more than 30 Mendelian human disorders [132]. An extended GGGGCC-repeat $[(G_4C_2)_n]$ within the C9orf72 gene is the most frequent monogenic cause of Frontotemporal Dementia and Amyotrophic Lateral Sclerosis c9FTD/ALS [133]. Similarly, accumulation of a CGG motif in the FMR1 gene underlies the Fragile X Syndrome, and is currently one of the most common identifiable genetic causes of mental retardation and autism [134]. In both repeat expansion disorders, recent evidence has suggested pronounced inter- and intraindividual repeat variability as well as focal changes in DNA methylation to modulate the disease phenotype [135–137]. Repeat expansions of up to 100-150 repeats can still be analyzed with conventional PCR or short read based approaches. However in the clinical diagnostic setting, 'analog' southern blot analysis are still state of the art to estimate the repeat length. Requiring multiple days and suffering from decreased resolution for longer repeats, these workflows would benefit from a sequencing based approach, both time- and resolution-wise. For applications in the research sector, the combination of a high counting accuracy and the single-molecule resolution of nanopore sequencing provides unforeseen potential to gain better insights. A SMRT sequencing is despite the higher accuracy of HiFi-reads less suitable due to the lacking sensitivity to DNA methylation on single molecule level.

To get a first impression of the visibility of STRs in nanopore sequenced samples, we used the publicly available data from the human GM12878 cell line. The inspection revealed a characteristic pattern in raw nanopore read signals spanning the c9orf72 STR locus (Fig. 5.1). To overcome current difficulties in characterizing expanded STRs we focused on three areas: i) optimization of nanopore sequencing and signal processing to capture STRs ii) development and implementation of a target enrichment strategy to increase efficiency and iii) integration of expansion

measurements with CpG methylation at the single molecule level. To enable a robust repeat analysis, we developed *STRique*, a general-purpose signal processing algorithm for the exact quantification of STR numbers in raw nanopore signals.

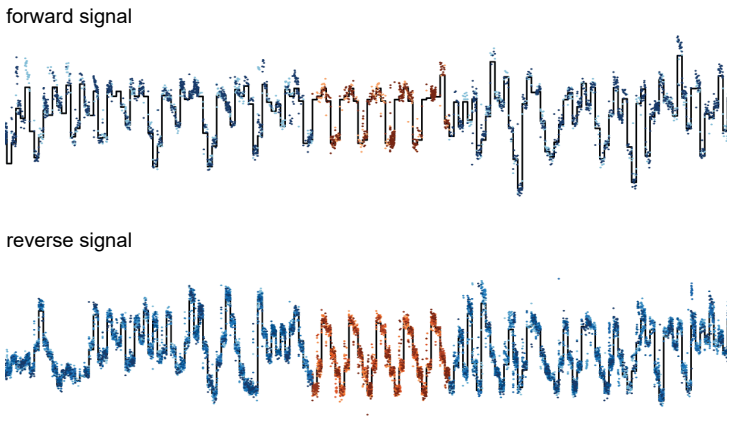


Fig. 5.1.: Multi signal HMM alignment of publicly available raw traces from two forward and eight reverse strand reads from the GM12878 cell line shows matching signal pattern in all reads [61]. Displayed are the current measurements as dots and the model signal as black line. Blue dots indicate current measurements identified as prefix or suffix sequence. Red dots indicate raw current measurements identified by *STRique* as belonging to the C9orf72-$(G_4C_2)_n$-STR. *STRique* detects in this case a $(G_4C_2)_5$-repeat.

5.2 Repeat quantification

Accurate counting of repeats with high resolution into even large expansion ranges serves as a first step during investigation of short tandem repeats. While a sequence based approach appears desirable due to a generally lower implementation complexity, the following sections illustrate the need for a signal based algorithm to exactly quantify the length of a short tandem repeat based on nanopore sequencing.

5.2.1 Sequence based repeat detection

To first benchmark existing repeat expansion evaluation methods we constructed, verified and nanopore sequenced plasmids with several synthetic $(G_4C_2)_n$-repeat lengths [138]. As a baseline, we manually counted repeats for a subset of reads, exploiting the clear visibility of the repetitive signal pattern (Fig. 5.1). Current (May 2019) production grade (*guppy* v3.0.3, high accuracy model) software developed by Oxford Nanopore Technologies (ONT) was used to translate the raw signal into the respective nucleotide sequence. In order to determine the repeat length with existing methods we deployed a decoy-alignment approach (Nanopore re-implementation of the STRetch algorithm [139]) and the *RepeatHMM* [140] package. For the alignment

method we added decoy-chromosomes to the reference genome, each with a distinct repeat length in the range of 3 to 100 (e.g. chr_9_3 to chr_9_100). The quantification is based on the assumption, that individual reads with possibly different repeat lengths will align to the decoy-chromosome with the best matching length. Resolution and range of this approach are limited by the set of additional reference sequences. *RepeatHMM* was initially developed and tested for tri- (SCA3/ATXN3) and pentanucleotide (SCA10/ATXN10) repeat expansions and originally only reports an estimated repeat length distribution per target locus The software was forked[1] and modified to also work with the c9orf72 hexanucleotide repeat and to provide individual counts per read.

Both methods were evaluated on our synthetic repeat sequences (8, 32, 50, 56 and 76 G_4C_2 repeats) and compared to the manual counted lengths. The analysis revealed, that current generation sequence based methods fail to satisfactorily resolve expanded STRs (Fig. 5.2).

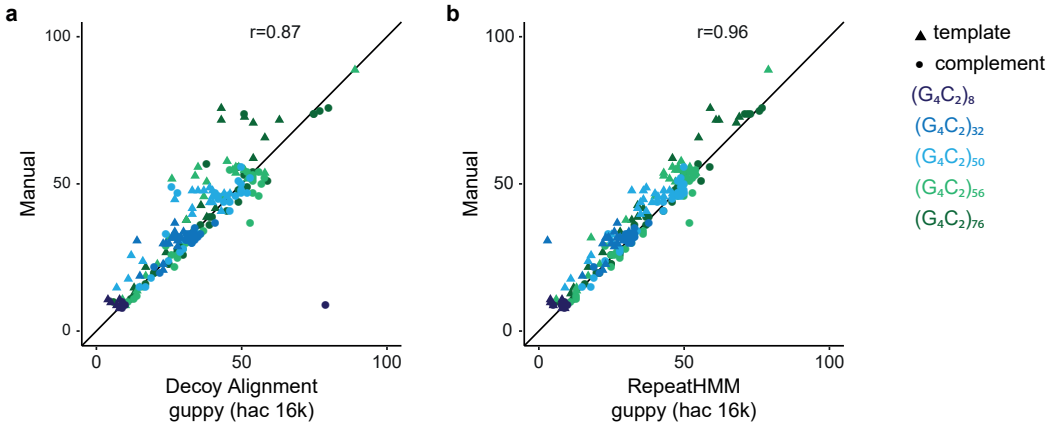


Fig. 5.2.: Manual counted set of plasmid reads on y-axis correlating with guppy basecalling and decoy alignment approach (**a**) and RepeatHMM (**b**) on x-axis. Only data points shown which could be evaluated with all three methods (n=15, 49, 45, 48, 47; Pearson correlation).

Appreciating the constant development and improvement of neural network based basecalling software [37] we systematically tested different versions and configurations. Specifically we deployed the previous state of the art software *albacore* and the research and technology demonstration tool *flappie* (both ONT). To keep memory requirements in a manageable range, these tools typically work on overlapping windows of the raw signal and combine these sub-sequences to the final output. For *guppy* and *albacore* the window size (default 1k and 10k respectively) is adjustable and we hypothesized that a larger basecalling window could improve the overall accuracy due to a larger context provided to the neural network. For *guppy* we further tested two different models for fast and high accuracy predictions and

[1] https://github.com/giesselmann/RepeatHMM, customized by Christian Rohrandt.

computed correlations with the manually obtained counts in Fig. 5.3 a. Whereas the deprecated *albacore* performed best in combination with the decoy alignment approach, sequences from *guppy* with high accuracy model and increased window size resulted in the highest correlation with manual counts when evaluated with *RepeatHMM* (Fig. 5.3 b). Further described and qualified in section 5.2.2, our method *STRique* outperforms any other approach on this level.

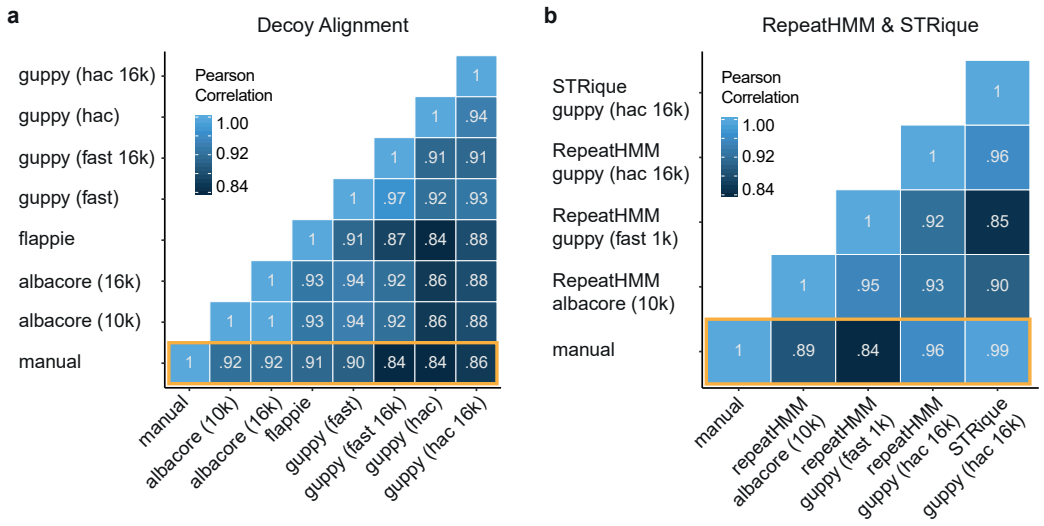


Fig. 5.3.: **a**, Correlation of manual counted repeat lengths with sequence base methods. Decoy alignment against reference with 3-100 repeats with Albacore (window 10k and 16k), Guppy (fast and hac mode, 1k and 16k window size) and Flappie basecalling (n=204 reads). **b**, Correlation of manual count with RepeatHMM and *STRique* results (n=204 reads).

To further assess the characteristics of existing workflows applied to larger repeat expansions, we next sequenced and analyzed the bacterial artificial chromosome (BAC) clone 239, generated from a c9FTD/ALS patient with an expected $(G_4C_2)_{800}$ repeat [141]. In absence of ground truth values per read we compare repeat counts across decoy alignment, *RepeatHMM* and *STRique* in Fig. 5.4. Mostly masked by a band of comparatively short repeats, only *STRique* is able to resolve a secondary peak at 800 repeats. More striking is the systematic strand bias observed in both sequence based methods resulting in generally more accurate counts for reads on the complement strand (GGGGCC repeat) compared to the template (GGCCCC repeat) strand.

In conclusion, we find that nanopore sequencing in general is capable of reading through expanded short tandem repeats. However, existing sequenced based methods fail to accurately quantify repeat lengths beyond ~32 hexanucleotide repeats. We further detect a strand and therefore sequence specific bias in the case of the c9orf72 G_4C_2 repeat, requiring re-evaluation of methods per target and worst case per software and model version.

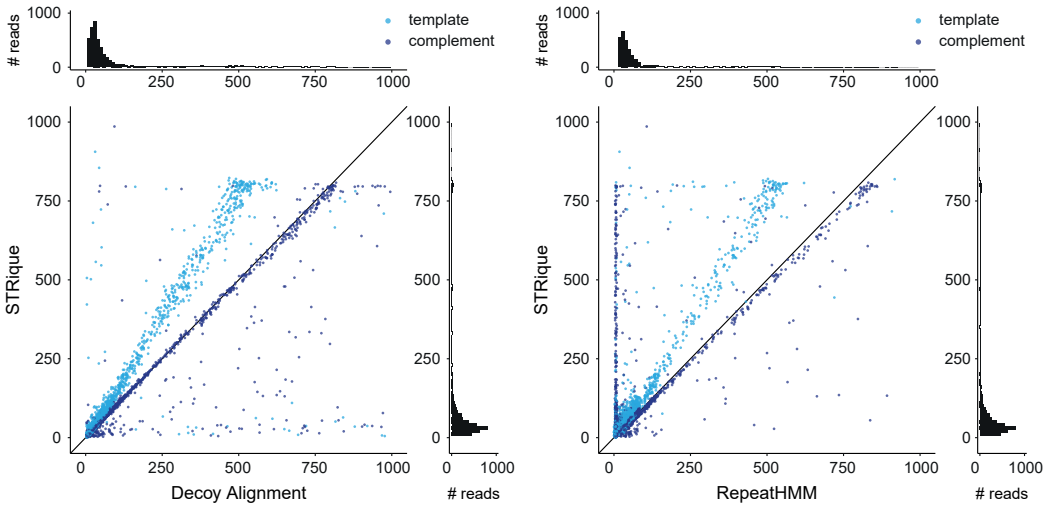


Fig. 5.4.: Comparison of repeat counts from *STRique*, decoy alignment based on guppy (high accuracy model, 16k window size) and repeatHMM based on guppy (high accuracy model, 16k window size) for BAC data. One dot (n=5004) per read passing all three approaches and colored by strand.

5.2.2 Signal based repeat detection

For overcoming inaccuracies of sequence based methods our *STRique* signal analysis software first identifies reads spanning a STR location by aligning the conventionally basecalled sequences to a reference [41]. Next, *STRique* maps the upstream and downstream boundaries of the repeat within each read more precisely with a signal alignment algorithm and, as a third step, quantifies the number of any given STR sequence with a Hidden Markov Model (HMM, Fig. 5.5).

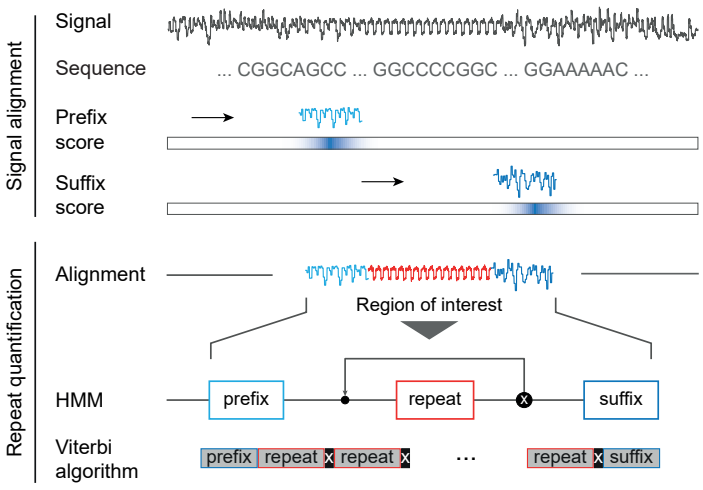


Fig. 5.5.: Repeat quantification enabled by raw signal alignment of flanking prefix and suffix regions and HMM-based count on the signal of interest.

Raw signal normalization is a crucial first step prior to the quantification process. Due to the reduced complexity of the repetitive signal segment, the overall distribution of measurements is skewed compared to reads from regular genomic contexts. Both, mean and median centered normalization fail to project a signal containing a STR expansion and being affected by offset, scaling and drift over time into a uniform range. We therefore applied a min-max normalization by scaling the 0.025 and 0.975 quantiles of the raw signal to the respective values of the pore model (cf. chapter 4.3).

With the normalized read and simulated signal fragments of the genomic sequence context around the STR, a signal alignment is used in order to locate prefix and suffix and to extract a region of interest. Within *STRique*, we use the distance based semi-global alignment described in section 4.4.1. Depending on the sequence complexity of prefix and suffix, a 100 to 150 bp frame around the repeat appeared to be robust. Reads with overlapping or reversed order of prefix and suffix alignments are flagged and dropped in this step. Furthermore the alignment scores, equivalent to the sum of absolute differences between simulated and raw signal are part of the *STRique* output and can be utilized to filter repeat counts in a post processing step.

Finally the actual repeat count is determined by a compound profile HMM composed of linear prefix and suffix modules, surrounding a single repeat instance with a feedback loop around (Fig. 5.5, section 4.4.2). The transition from repeat end to repeat begin enables the HMM to stay in the repeat associated states to model arbitrary long repeat expansions. The Viterbi algorithm assigns the most likely state from prefix, repeat and suffix to each observation of the normalized signal given the models states and transitions. The number of passes through the feedback transition is reported as the detected repeat count, the precise positions of prefix end and suffix begin are provided to enable masking of the repetitive signal fragment for downstream processing (cf. section 5.3.1).

Aggregated *STRique* repeat counts matched closely gel electrophoresis profiles (Bioanalyzer) from our synthetic repeat constructs and could be confirmed on the single molecule level by manually counting repeat patterns in raw signal traces (Fig. 5.6).

STRique is written in python3 with a C++ extension for the signal alignment, supports multiprocessing and is automatically build and tested using Travis-CI. *STRique* is deployed either as a python3 virtual environment or as a standalone Docker container.

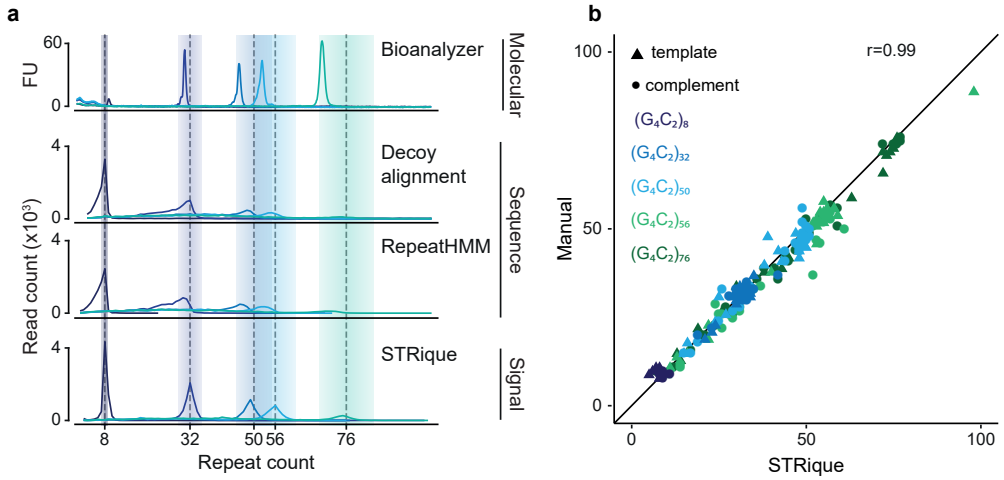


Fig. 5.6.: **a**, Bioanalyzer electropherogram, decoy alignment, *RepeatHMM* and *STRique* counts of synthetic $(G_4C_2)_n$ repeats. **b**, Manual counted set of plasmid reads on y-axis correlating with *STRique* raw signal pipeline on x-axis. Only data points shown which could also be evaluated with methods in Fig. 5.2 (n=15, 49, 45, 48, 47; Pearson correlation).

5.2.3 Repeat expansion in C9orf72 and FMR1

The nanopore sequencing of repeats integrated into plasmid or BAC backbones typically yields thousands of reads containing the expanded repeat. For the application in disease contexts, the sequencing of DNA from e.g. patient-derived induced pluripotent stem cell lines is favored, as it for instance preserves the epigenetic landscape around the repeat. However, a straightforward whole genome sequencing would only yield few reads covering the repeat of interest. From e.g. 30 Gbp throughput, equivalent to 10 fold genome wide coverage, on average only 10 reads would be expected to cover any target repeat. For the monoallelic expansions in the C9orf72 and FMR1 genes, only half of those would be informative to infer the repeat length.

To increase the amount of reads covering any repeat expansion of interest, we therefore set up a CRISPR-Cas9 target enrichment strategy. Briefly, during library preparation, the method cuts the DNA on defined guide sequences next to the target locus. Following on a dephosphorylation step during fragmentation, only the Cas9 cleaved and therefore phosphorylated ends are capable to ligate the ONT sequencing adapter. Hence only fragments starting on definable guide sequences are sequenced on the nanopore flow cell. We applied enrichment and quantification to one C9orf72 repeat expansion patient derived stem cell line (24/5#2) and two patient derived cell lines (SC105-iPS6/iPS7) from another patient with a FMR1 repeat expansion (Fig. 5.7).

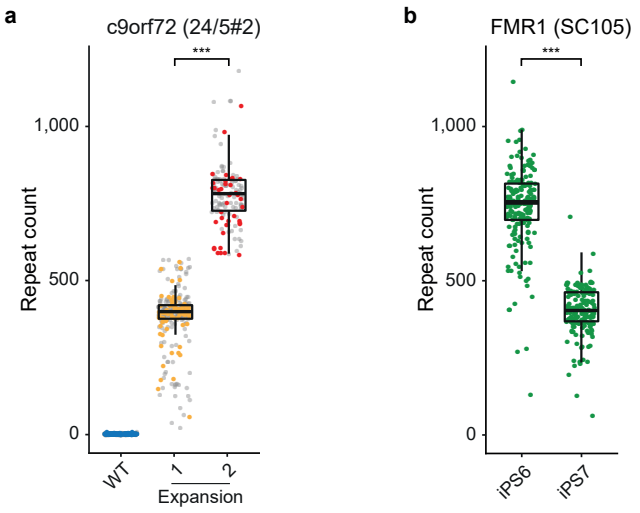


Fig. 5.7.: **a**, Repeat quantification of sample 24/5#2 at the C9orf72 locus, revealing two distinct repeat bands of ~450 and ~750 G_4C_2 repeats (n = 1,810, 738 and 363 evaluated reads with a difference in repeat length of 392 (95% confidence interval (CI): 383 to 400), $P < 2.2 \cdot 10^{-16}$). Colored points indicate reads used in Fig. 5.9b. WT, wild-type. **b**, Repeat quantification of the SC105iPS6 and SC105iPS7 samples at the FMR1 locus (n = 174 and 168 evaluated reads with a difference in repeat length of -343 (95% CI: -361 to -325), $P < 2.2 \cdot 10^{-16}$). P values in a and b were obtained by a two-sided Wilcoxon rank-sum test; $***P < 0.001$. Data are presented as boxplots (centerline, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range).

In the C9orf72 case, the analysis revealed two distinct repeat expansion bands next to the wild-type allele with ~450 and ~750 G_4C_2 repeats respectively. Even though derived from the same male patient, the two FMR1 samples showed different repeat lengths with a difference of 343 CGG repeats. During development and improvement of the enrichment protocol, both samples were sequenced multiple times on the MinION and PromethION platform. Nonetheless, the two repeat expansion cluster are already detectable from the output of a single MinION sequencing run with up to thousand reads on target (Fig. 5.8, FAK67994)

In summary we show, that even heterogeneous expansion distributions of any short tandem repeat can be sequenced and quantified from a single MinION flow cell. In comparison to established southern blotting, our approach increases the accuracy and adds the resolution of single molecule counts. In combination with target enrichment, the workflow could be deployed to provide same-day analysis in a clinical environment.

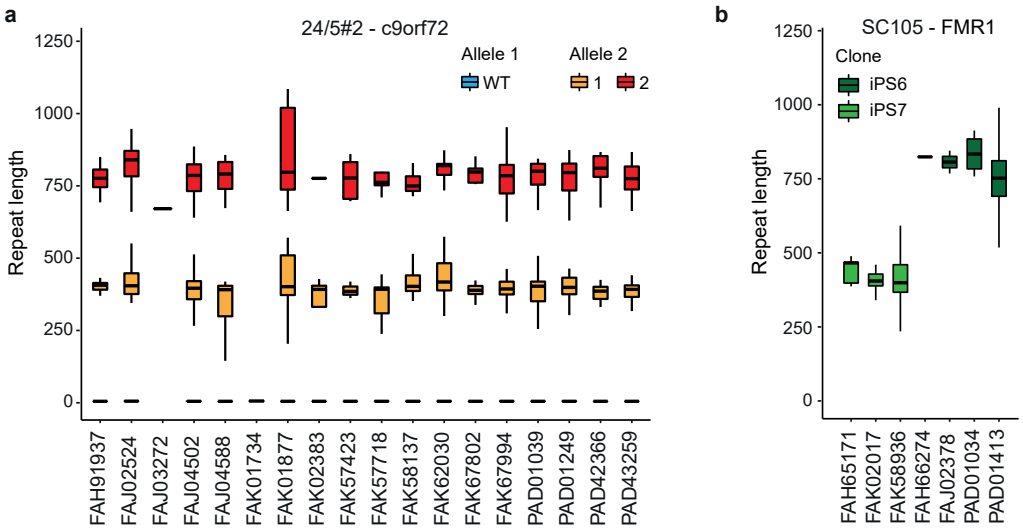


Fig. 5.8.: **a**, C9orf72 target enrichment flow cells for patient 24/5#2 **b**, FMR1 enrichment flow cells of SC105iPS6/iPS7. (FA*: MinION, PAD*: PromethION, number of reads per boxplot are in Supplementary Tables B.1 and B.2). Data presented as boxplots (centerline, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range; outliers not shown)

5.3 Methylation detection

The epigenetic modification of C9orf72 and FMR1 loci have been correlated with STR expansion status and patient characteristics in both disorders, however without quantification at the single molecule level so far [137, 142]. Additionally, in c9FTD/ALS patients pervasive CpG methylation of the G_4C_2-repeat itself has been reported [143]. Assessed with a strictly qualitative assay, the expanded STR itself was reported to be methylated in the majority of cases examined. A similar observation has been directly implicated in the pathogenesis of FXS, where a CGG repeat expansion at the FMR1-locus beyond a threshold of > 200 repeats leads in most cases to the silencing of the entire FMR1-gene through CpG-methylation [144]. Exploiting the visibility of 5-methylcytosine in nanopore sequencing, we extend *STRique* in the following sections to integrate repeat length and DNA methylation status on the same molecule.

5.3.1 Region methylation detection

Nanopore reads spanning an expanded STR location can be aligned to a reference genome using standard tools like *minimap2* [41]. Though, due to the potentially large insertion, the reads are commonly soft-clipped from one side up to the STR, mapping either prefix or suffix. While this is sufficient to identify reads spanning the

target locus during quantification of the repeat length, it does not allow the direct readout of the DNA methylation status on both sides of the repeat. We therefore use the signal level information of the exact position of the STR within the read from the previous quantification and provide a script (fast5Masker.py) to trim the STR signal fragment from each read (cf. red signal segment in Fig 5.5). The resulting traces are stored as regular fast5 files and can be processed with any standard pipeline. The manipulation of the raw signal leads to minor basecalling errors directly adjacent to the STR, but enables the mapping of both flanking sequences to the reference genome.

Hereafter we integrated single read CpG methylation analysis of regions adjacent to the c9orf72 STR using *nanopolish* [102] with our *STRique* results (Fig. 5.9 a). We found that in the 24/5#2 line all reads with STR expansions > 750 repeats showed a significantly increased methylation level at the promoter CpG island. In contrast all wild-type reads and those with ~450 repeats were not or only partially methylated (Two sided Wilcoxon rank sum test p < 0.001, Fig. 5.9 b).

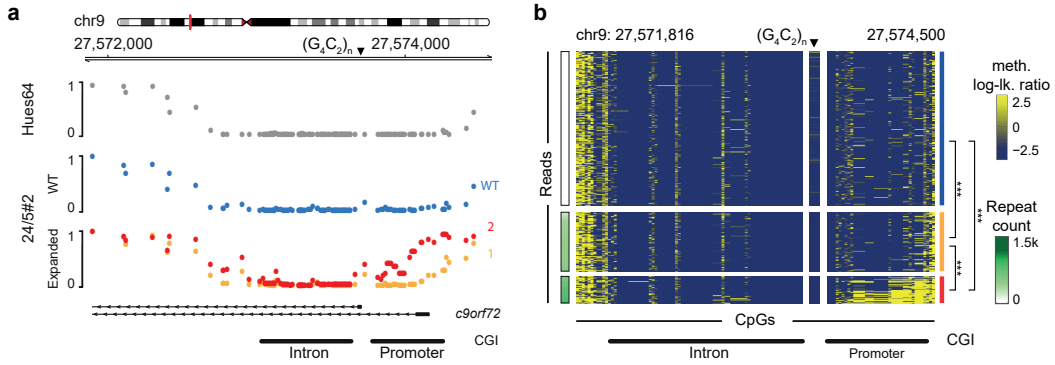


Fig. 5.9.: **a**, C9orf72 methylation status in HUES64 as measured by whole-genome bisulfite sequencing. The wild-type (blue) allele and expanded (ex; orange) alleles (with 450 and 750 $(G_4C_2)_n$ repeats (red), respectively) are shown for patient 24/5#2, as measured by nanopore sequencing. **b**, Single read nanopore methylation of C9orf72 covering reads from the minus strand (n = 259, 100 and 43 rows per block) sorted by detected repeat length (rows, single read; columns, single CpGs). CpGs with logP ratio > 2.5 are considered methylated, while those with logP ratio < -2.5 are considered unmethylated. The median methylation difference (95% CI) and P value (determined by two-sided Wilcoxon rank-sum test on mean promoter CGI methylation) for comparisons were as follows: $WT - ex450 : 3.9 \cdot 10^{-5}(4.8 \cdot 10^{-6} to 3.4 \cdot 10^{-2}), P = 5.3 \cdot 10^{-9}; WT - ex750 : 0.56(0.46 - 0.64), P < 2.2 \cdot 10^{-16}; ex450 - ex750 : 0.53(0.40 - 0.64), P < 2.2 \cdot 10^{-16}; ***P < 0.001$.

Masking segments of the raw nanopore signal is a considerable change of the original measurement. We therefore validated the region methylation detection approach on our previously introduced BAC data. Amplified in bacteria, the artificial chromosome is free of CpG methylation but contains a larger sequence window around the repeat on chromosome 9 making it comparable to the patient data. Applying the same workflow of repeat quantification, masking and nanopolish methylation detection,

we find no 5-methylcytosine signal in any of the BAC reads independent of the repeat length (Fig. 5.10 a,b). Comparing the mean methylation rates between patient 24/5#2 and BAC, we find no evidence for a difference on the intron CGI, but significant level differences in particular for the ~750 repeat cluster on the promoter CGI.

In summary, the STR detection on signal level facilitates the reconstruction of wild-type like reads from large STR expansions, enabling execution of established workflows.

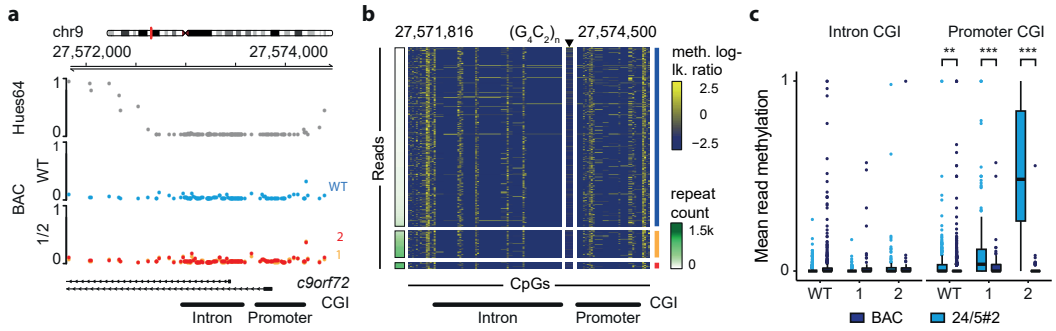


Fig. 5.10.: **a**, Methylation status of c9orf72 region in BAC data for repeats < 200 (WT), 200-750 (Cluster1,orange) and > 750 (Cluster2,red) and control (HUES64, WGBS) **b**, Single read methylation on a sample of 500 BAC minus strand reads sorted by repeat count (row split 200 and 750 repeats, n=423,63,14). **c**, Difference in mean CGI methylation of intron and promoter per read on minus strand. Reads binned by detected repeat length for BAC (n=2066 WT; 315 Cluster1; 72 Cluster2) and patient 24/5#2 (n=925 WT; 362 Cluster1; 153 Cluster2). Two sided Wilcoxon rank sum test, corrected for multiple testing (Holm), q-vals: * 0.05 - 0.01; ** 0.01 - 0.001; *** < 0.001. Median methylation differences between promoter CGI [95%CI] for WT $-2.3e^{-5}[CI: -5.6e^{-6}: -1.5e^{-5}, q = 7.4e^{-3}]$ and Cluster1 $-0.01[CI: -7.1e^{-5}: -3.4e^{-2}, q = 1.4e^{-17}]$ and Cluster2 $-0.46[CI: -0.58: -0.37, q = 1.0e^{-26}]$.

5.3.2 Repeat methylation detection

Due to the intrinsic heterogeneity in STR length, reference genome based methods such as *nanopolish* cannot be used to determine CpG methylation on the repeat expansion itself. To detect 5mC modifications on STRs, we extended *STRique* by employing a parallel HMM with unmodified- and 5mC-paths. Taking the expected signal levels for native and 5mC modified DNA from nanopolish, the profile-HMM block modeling the repeat is replaced by two parallel blocks, allowing the full HMM to switch between methylated and native path during each repeat iteration. This single read analysis returns a methylation state for each tandem repeat, which then can be summarized into the mean repeat methylation level over the whole repetitive sequence.

When applying the methylation-aware *STRique*, all expanded FMR1-STRs in nanopore reads from patient SC105 are found to be highly methylated (Fig. 5.11 a), consistent with previous analyses [142]. We next evaluated this approach on plasmids containing n=76 synthetic G_4C_2 and n=99 CGG-repeats (Addgene, #63089), which were covalently modified with the methyltransferase M.SssI (Fig. 5.11 b,c). In addition we tested the algorithm on $(G_4C_2)_n$-containing reads from patient-derived DNA, which had been modified with M.SssI in vitro. In summary, we found that *STRique* can determine the repeat CpG methylation state correctly in all positive and negative controls evaluated. Surprisingly though, all reads covering the C9orf72-STR from our patient-derived samples showed little to no CpG-methylation, independently of the repeat expansion length or methylation status of the promoter CGI (Fig. 5.9 b, center column).

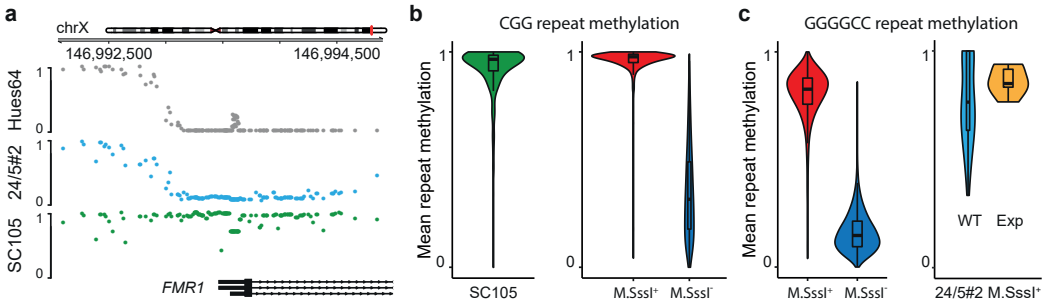


Fig. 5.11.: **a**, FMR1 region methylation in SC105iPS6/iPS7 compared to HUES64 WGBS and patient sample 24/5#2. **b**, CGG mean repeat methylation status detected by *STRique* for SC105 (n=197) and synthetic plasmid control with 99 repeats treated with $M.SssI^{+/-}$ (5mC level on minus strand, n=1232 $M.SssI^{+}$; n=11991 $M.SssI^{-}$). **c**, GGGGCC repeat methylation status for plasmid control with 76 repeats treated with $M.SssI^{+/-}$ (n=2939 $M.SssI^{+}$; n=31280 $M.SssI^{-}$) and patient sample 24/5#2 treated with $M.SssI^{+}$ (5mC level on minus strand, n=52 WT and n=6 Cluster1). Data in (b-c) presented as violin plots with overlayed boxplots (centerline, median; box limits, first and third quartiles; whiskers, 1.5x interquartile range; outliers not shown).

5.4 Summary

Our results demonstrate the precise and multi-layered molecular characterization of pathological short tandem repeat expansions. The CRISPR/Cas-nuclease-target enrichment and *STRique* can be rapidly adapted to any other genomic region of interest, ensuring broad applicability to overcome challenges associated with the single molecule analysis. This allows for immediate integration of genetic and epigenetic signals associated with unstable repeat expansions or any other as of yet unsequenceable genomic regions in human health and disease. This type of analysis improves diagnostic workflows in regard to accuracy and resolution of unstable

repeat expansion while enabling efforts to gain mechanistic insights into effects on differentiation, aging and future therapeutic agents that modify DNA methylation.

6 Discussion

But could you not also do that with short reads? - Expressing an initial scepticism against a novel technology, nanopore sequencing is, six years after the first release in 2014, still termed the emerging technology. Next-generation sequencing remains the established and widely undisputed reference method. However, after significant improvements, addressing shortcomings of throughput and accuracy, both third-generation platforms have evolved to become the primary technology for genome assembly and are commonly used for the analysis of viral and bacterial genomes. Yet, a broader replacement of NGS methods is not in sight, raising the question, why the introductory question is not more frequently re-phrased to: *Would nanopore here not be the more suitable platform?* Or: *Would you not confirm this by using nanopore sequencing?*

The underlying research question of this work is therefore to analyze unique use-cases for nanopore sequencing, identify key limitations and provide a comprehensive investigation of the current status. Despite many improvements, the application of nanopore sequencing in the context of large mammalian and plant genomes remains challenging. At the same time are the benefits of long reads known: Reliable alignments in repetitive regions, no duplicates and biases by library amplification and the linkage of distant genetic and epigenetic features on single molecule level. Yet, the stable throughput, higher read accuracy and established workflows of NGS technologies justify the commitment to nanopore sequencing only for applications infeasible with short reads.

A niche to gain traction at applying nanopore sequencing is the analysis of repetitive regions in the human genome. For comparison, using short reads, the reachable part of a repeat is limited by the read length, leading to ambiguous alignments, once a read contains only repeat sequence. Short tandem repeats (STRs) are accumulations of three to six nucleotide long sequence patterns, in disease cases expanded to hundreds of copies. The epigenetic analysis of repeats is a unique feature of nanopore sequencing, considering that SMRT sequencing is not sufficiently sensitive to 5-methylcytosine on single-molecule level. Repeat detection by sequencing is the digital advancement over southern blot based quantification, increasing the accuracy

and reducing the turnaround time. *STRique* is a bioinformatic analysis software, developed to integrate the quantification of short tandem repeats with their methylation state on individual read level. The algorithmic key features of *STRique* are raw signal alignment and the sequence to signal annotation. Due to the oversampling of the sequencer, raw signal traces are roughly ten times longer compared to their corresponding sequence and, with respect to noise and time warping, generally more difficult to handle. Yet, only a signal based counting algorithm allows to bypass systematic errors induced during conventional basecalling of tandem repeats. Hidden Markov Models (HMM) are a powerful resource to align nanopore signals to a reference sequence, with the most prominent usage in the *Nanopolish* package. However, their computational complexity is scaling quadratically with the number of hidden states, limiting the usage to restricted signal and sequence windows of interest. The purpose of the signal alignment is therefore, to extract a signal segment covering the repeat with sufficient flanking sequence to anchor a counting HMM. The flexibility of the HMM is needed over a static alignment, since the repeat length is unknown and varying between reads. A looping transition around a single repeat profile allows the model to iterate through an arbitrary number of repeat patterns. The signal alignment is the enabling and at the same time limiting step: The Viterbi path through the HMM is the most likely state-sequence given the observed signal, fed with a wrong signal alignment, the model will still yield a repeat count. A filtering based on signal alignment scores is therefore crucial to exclude low quality counts. A useful side-effect of the signal alignment is the ability to mask regions within the read. *STRique* provides a script to cut out the expanded repeat from the raw signal and creates wild-type like reads passing any conventional downstream pipeline, to detect for instance DNA-methylation in the repeat-flanking sequence. Taken together, *STRique* is an example for a unique nanopore application, enabling more detailed investigations of short tandem repeat expansions.

Bioinformatics for nanopore sequencing are undergoing constant development. Even a stable nanopore workflow needs, in contrast to NGS analysis, constant maintenance to factor in accuracy improvements, new pore generations, but also for example recent file format and compression method changes. Zooming out from individual tool- to pipeline-level, the streamlined processing and data archiving has received very little attention in the nanopore community. Computational setups ranging from few MinION flow cells analyzed on a Laptop to high-throughput PromethION sequencing in combination with GPU accelerated cluster environments are complicating, if not inhibiting, a uniform processing strategy. The motivation behind the development of the *Nanopype* pipeline is therefore not only to wrap the processing, but also set up a scaling data management. While certainly site specific

on the implementation detail, the concepts of the pipeline backend transfer to any institute aiming to set up a nanopore workflow. The ONT device control software *MinKnow* is providing raw data in batches of currently 4k reads packed into *fast5* files. Further transfer to permanent storage, indexing and archiving are essential for a high-throughput environment. The organization in batches is, as opposed to a single *fastq* file per NGS run, beneficial for distributed processing. *Nanopype* operates on batches whenever possible, submitting for instance alignments in batches bears the advantage that multiple compute nodes can process a single sequencing run and that only few reads need to be re-processed in case of errors. An abstraction layer around the raw data handling is needed to be robust against different file formats (single- and bulk-*fast5*) and changing output directory structures of the *MinKNOW* software.

Using workflow management systems such as *Nextflow* or *Snakemake* over plain shell script implementations greatly simplifies the development and maintenance of a pipeline. *Snakemake* workflows are becoming more popular for both, standalone pipeline development, but also to enhance the reproducibility of publications by wrapping the applied toolchain into a workflow. Other pipelines in the field are mostly linear in a way, that they utilize a single tool per processing step and provide specialized outputs to analyze e.g. isoforms or single-nucleotide polymorphisms (SNPs). Tailored to the requirements of methods development, *Nanopype* provides basic processing with interchangeable methods for each module, supporting for instance *minimap2*, *NGMLR* and *GraphMap* alignments. Among the greatest advantages of using *Snakemake* is the support of cluster schedulers and the build-in support for *Singularity* containers. With regard to increasing throughput and fast paced algorithmic accuracy improvements, distributed processing of nanopore datasets is an essential ability to operate on a competitive level. The deployment of pipeline modules as *Singularity* containers is of less importance for the local usage, but is setting a new state-of-the-art for providing exactly reproducible workflows. *Nanopype* is unique in not only embedding the order and parameters of applied tools, but also enforcing their version. Tool installations are commonly left to the user, by either expecting to find required tools in the systems *PATH* variable or by providing options with full path configurations. Both ways allow the execution of the workflow, but will produce divergent outputs depending on installed versions. Within *Nanopype*, installation wrappers ensure that tool versions are tied to the pipeline version, guaranteeing consistent outputs across labs and experiments. Lastly, after building the *Singularity* images as part of the deployment process, functionality tests can be executed with little overhead on minimal example datasets. The broad functional range of *Nanopype* is ideal for basic processing and serves as a baseline for

both, users and developers. At the same time is the all-in-one approach increasingly complex to maintain and extend. An example is the trend to apply deep learning algorithms to raw nanopore signals, not only for basecalling, but also for barcode demultiplexing, variant- or base-modification-detection. Popular neural network frameworks are *Tensorflow* (v1 and v2) and *Pytorch*, which have to be provided by the pipeline. The current implementation would for instance not allow to have parallel applications requiring different versions of *Tensorflow*. A subject of future work is therefore the further isolation of included tools, for instance through distinct python virtual environments for compatible applications. In order to not further increase the complexity of the core pipeline, the *Snakemake* subworkflow feature could be used to implement nested extension pipelines, each requiring *Nanopype* outputs as external inputs.

Nanopore sequencing: What's coming next?

The exploration of established use-cases in the literature review reveals the broad applicability of third-generation sequencing for the analysis of bacterial and viral genomes. Genome assembly and structural variant detection are prominent examples for applications, where NGS sequencing is no longer the primary technology. The thought experiment of what would be needed to fully replace NGS by nanopore sequencing allows to assess remaining limitations. If nanopore reads had perfect accuracy, not only on sequence, but also on base-modification level, there would be no further need to store large raw datasets for future re-processing. A higher accuracy would also reduce the required coverage, facilitating the widespread sequencing of larger genomes with present throughput. On the other hand are short reads convenient for count-based methods such as ChIP-Seq, where long reads would impede peak-calling. Viewing third-generation sequencing as an extension, rather than a potential replacement, appears therefore to be the most appropriate perception.

From a hardware and wet-lab perspective, shrinking the sequencing devices using voltage sensors or solid state pores could help to increase throughput and at the same time decrease the amount of input material needed for library preparation. From a software developer perspective, increasing the sequence and base modfication detection accuracy on existing data would have wide-ranging impact. Processing infrastructure and signal segmentation methods of this work are the foundation for future work on basecalling and methylation detection algorithms. Novel transformer neural network architectures have demonstrated impressive results on translation and speech to text conversions. Translating an event table of feature vectors into

a genomic sequence is from an algorithmic point of view very similar and the performance of attention based learning over classical recurrent layers will be exciting to investigate.

A Nanopype Supplement

A.1 Listings

```
rule minimap2:
    output:
        bin = "bin/minimap2"
    threads: config['threads_build']
    shell:
        """
        mkdir -p src && cd src
        if [ ! -d minimap2 ]; then
            git clone https://github.com/lh3/minimap2 --branch v2.14 --depth=1 && cd minimap2
        else
            cd minimap2 && git fetch --all --tags --prune && git checkout tags/v2.14
        fi
        make clean && make -j{threads}
        cp minimap2 ../../{output.bin}
        """
```

Listing A.1: Snakemake installation wrapper

Sequencing Report

V65_v2

3	Flow cells
121.1 GB	Total bases
29.0 kB	N50
1763.54 GB	Total disk usage

Nanopype v1.0.1-11-ga124850

1 Flow cells

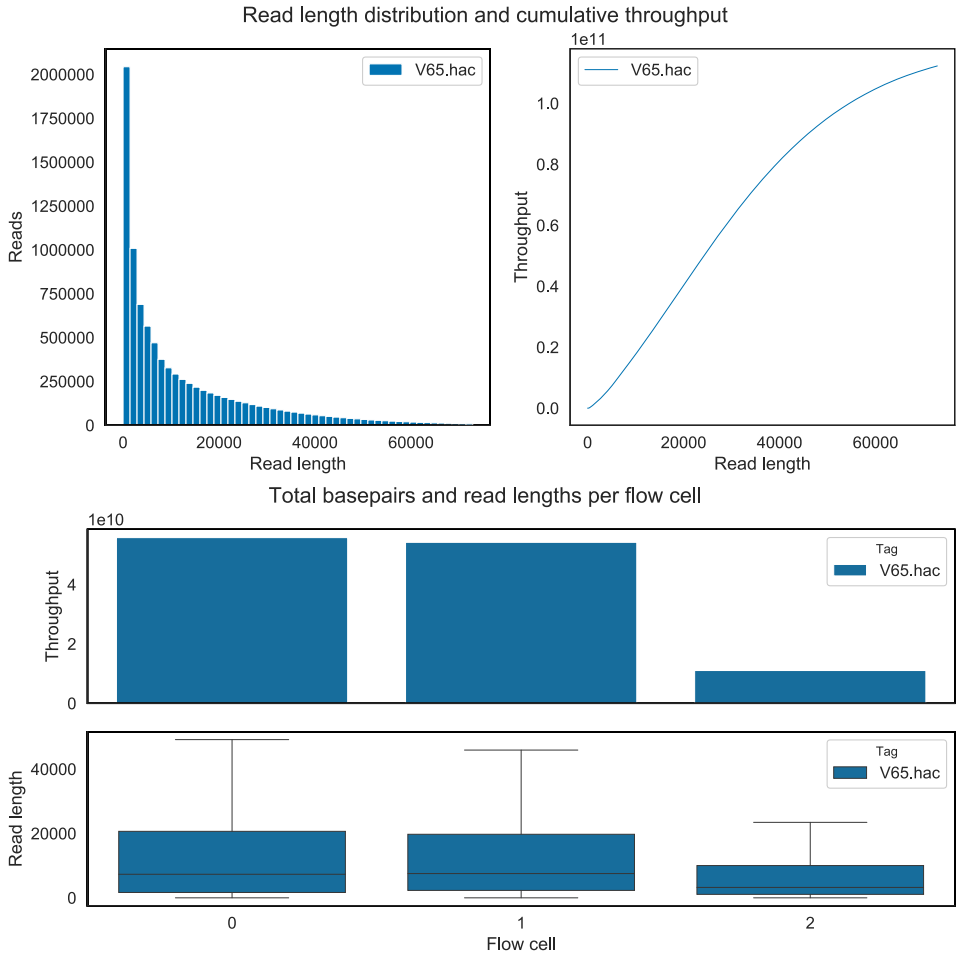
0: 20200708_PAE91419_FLO-PRO002_SQK-LSK109_V65_WT
1: 20200709_PAE86216_FLO-PRO002_SQK-LSK109_V65_WT
2: 20200710_FAN41394_FLO-MIN106_SQK-LSK109_V65_WT

2 Basecalling

2.1 Summary

		Sum	Mean	Median	N50	Maximum
V65.hac	guppy	121110031025	13239	6603	29048	1315417

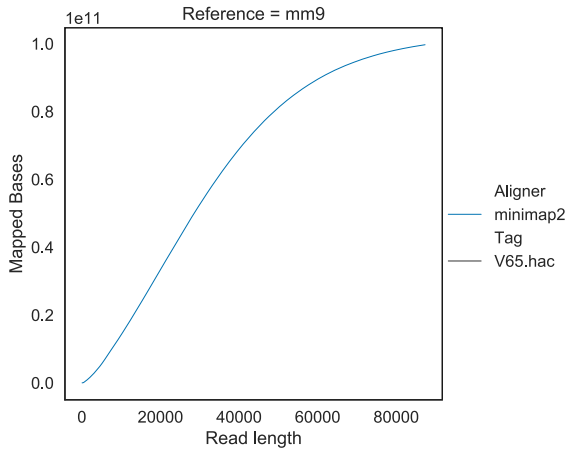
2.2 Guppy:



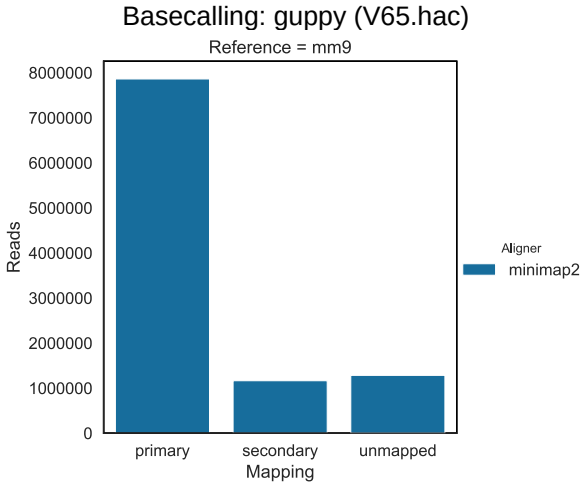
3 Alignments

3.1 Mapped bases (primary alignments)

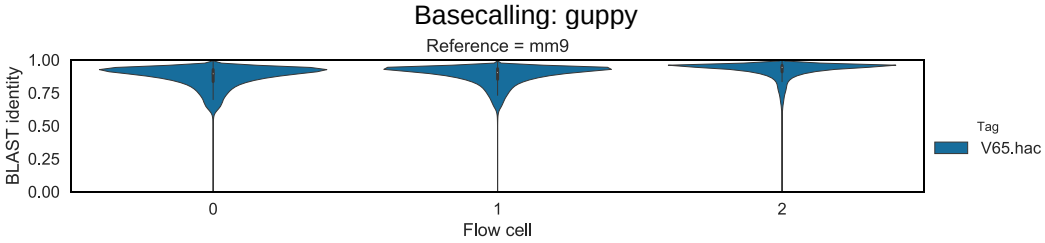
Basecalling: guppy



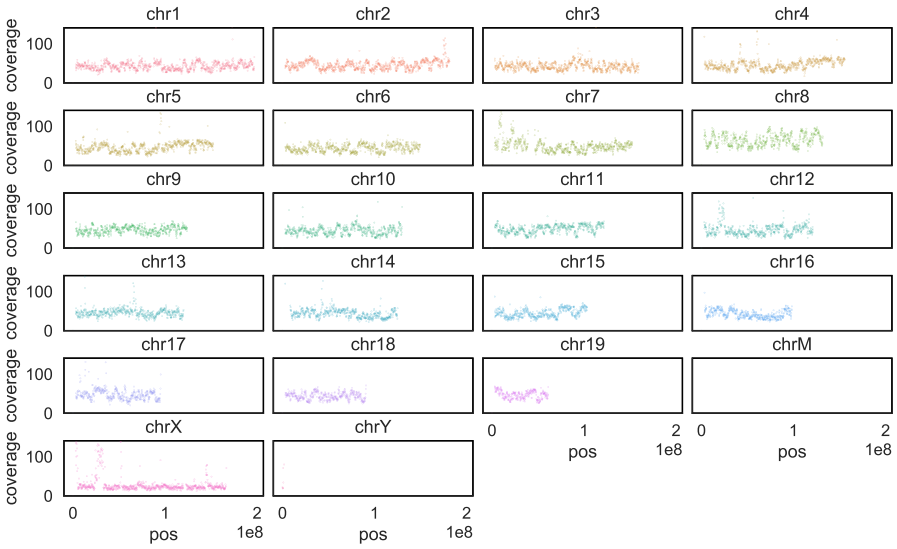
3.2 Mapped reads



3.3 Read identity (primary alignments, all aligners)

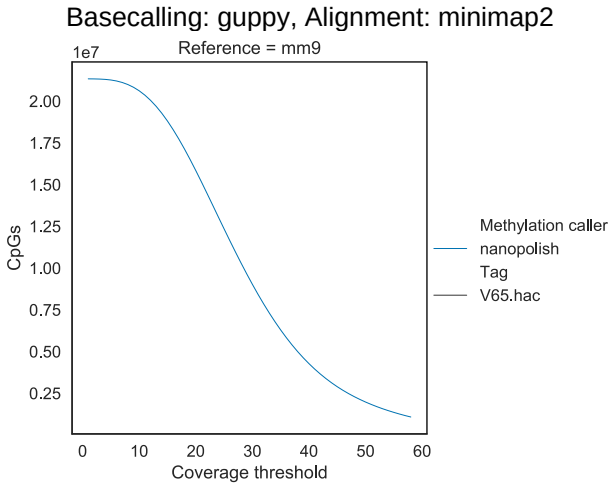


3.4 Coverage



4 Methylation

4.1 CpG coverage



B

STRique Supplement

Tab. B.1.: Cas12 target enrichment throughput per flow cell

run	reads	valid	wild type	expanded	target	sample
FAH91937	74	67	36	31	C9orf72	24/5#2
FAH91937	5	3	3	NA	FMR1	24/5#2
FAJ02524	139	87	51	36	C9orf72	24/5#2
FAJ02524	2	0	0	NA	FMR1	24/5#2
FAJ03272	14	1	0	1	C9orf72	24/5#2
FAK02383	10	7	2	5	C9orf72	24/5#2
FAK02383	5	5	5	NA	FMR1	24/5#2
FAK02402	1	1	1	0	C9orf72	24/5#2
FAK57423	22	18	9	9	C9orf72	24/5#2
FAK57423	25	22	22	NA	FMR1	24/5#2
FAK67802	102	93	53	40	C9orf72	24/5#2
FAK67802	63	55	55	NA	FMR1	24/5#2
PAD01039	228	122	72	50	C9orf72	24/5#2
PAD01039	19	10	10	NA	FMR1	24/5#2
PAD01249	601	436	290	146	C9orf72	24/5#2
PAD01249	623	491	491	NA	FMR1	24/5#2
PAD42366	152	122	69	53	C9orf72	24/5#2
PAD42366	102	90	90	NA	FMR1	24/5#2
FAH66294	28	21	21	NA	C9orf72	iPS6
FAH66294	1	0	0	0	FMR1	iPS6
FAJ02378	350	291	291	NA	C9orf72	iPS6
FAJ02378	4	2	0	2	FMR1	iPS6
PAD01034	151	96	65	NA	C9orf72	iPS6
PAD01034	5	4	0	4	FMR1	iPS6
PAD01413	779	565	564	NA	C9orf72	iPS6
PAD01413	317	166	0	166	FMR1	iPS6
FAK02017	2	0	0	NA	C9orf72	iPS7
FAK02017	28	11	0	11	FMR1	iPS7
FAK58936	283	261	261	NA	C9orf72	iPS7

FAK58936	181	147	0	147	FMR1	iPS7

Tab. B.2.: Cas9 target enrichment throughput per flow cell

run	reads	valid	wild type	expanded	target	sample
FAJ04502	639	463	284	179	C9orf72	24/5#2
FAJ04502	316	233	233	NA	FMR1	24/5#2
FAJ04588	94	67	38	29	C9orf72	24/5#2
FAJ04588	46	39	39	NA	FMR1	24/5#2
FAK01734	2	2	2	0	C9orf72	24/5#2
FAK01734	1	1	1	NA	FMR1	24/5#2
FAK01877	36	31	17	14	C9orf72	24/5#2
FAK01877	56	47	47	NA	FMR1	24/5#2
FAK57718	73	60	37	23	C9orf72	24/5#2
FAK57718	69	59	59	NA	FMR1	24/5#2
FAK58137	157	128	75	53	C9orf72	24/5#2
FAK58137	127	109	109	NA	FMR1	24/5#2
FAK62030	151	126	82	44	C9orf72	24/5#2
FAK62030	100	79	79	NA	FMR1	24/5#2
FAK67994	1041	884	575	309	C9orf72	24/5#2
FAK67994	367	319	319	NA	FMR1	24/5#2
PAD43259	364	298	187	111	C9orf72	24/5#2
PAD43259	197	154	154	NA	FMR1	24/5#2

Bibliography

1. Zymo research. *What is Epigenetics | ZYMO RESEARCH* `https://www.zymoresearch.de/pages/what-is-epigenetics`. Online; accessed 19-10-2020. 2020.

2. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20.** Number: 4 Publisher: Nature Publishing Group, 207–220 (Apr. 2019).

3. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21.** Number: 3 Publisher: Nature Publishing Group, 381–395 (Mar. 2011).

4. Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics* **32,** 42–56 (Jan. 2016).

5. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* **20.** Number: 10 Publisher: Nature Publishing Group, 590–607 (Oct. 2019).

6. Holliday, R. & Grigg, G. W. DNA methylation and mutation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. Special Issue In Memory of Max Clark, a Pioneer in Fundamental Mutation Reserach* **285,** 61–67 (Jan. 1993).

7. Bird, A. DNA methylation patterns and epigenetic memory. *Genes & Development* **16.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 6–21 (Jan. 2002).

8. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13,** 1095–1107 (Aug. 1992).

9. Capper, D., Jones, D. T. W., Sill, M., *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555.** Number: 7697 Publisher: Nature Publishing Group, 469–474 (Mar. 2018).

10. McGuire, A. L., Gabriel, S., Tishkoff, S. A., *et al.* The road ahead in genetics and genomics. *Nature Reviews Genetics* **21.** Number: 10 Publisher: Nature Publishing Group, 581–596 (Oct. 2020).

11. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74.** Publisher: National Academy of Sciences Section: Biological Sciences: Biochemistry, 5463–5467 (Dec. 1977).

12. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* **30,** 418–426 (Sept. 2014).

13. Illumina. *An introduction to Next-Generation Sequencing Technology* `https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`. Illumina.

14. Frommer, M., McDonald, L. E., Millar, D. S., *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* **89.** Publisher: National Academy of Sciences Section: Research Article, 1827–1831 (Mar. 1992).

15. Ziller, M. J., Hansen, K. D., Meissner, A. & Aryee, M. J. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods* **12.** Number: 3 Publisher: Nature Publishing Group, 230–232 (Mar. 2015).

16. Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34,** 666–681 (Sept. 2018).

17. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* **34,** 518–524 (May 2016).

18. Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. *S2ORC: The Semantic Scholar Open Research Corpus* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, July 2020), 4969–4983.

19. Peroni, S. & Shotton, D. OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies* **1.** Publisher: MIT Press, 428–444 (Jan. 2020).

20. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *arXiv:1607.00653 [cs, stat]*. arXiv: 1607.00653 (July 2016).

21. Perozzi, B., Al-Rfou, R. & Skiena, S. *DeepWalk: online learning of social representations* in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (Association for Computing Machinery, New York, NY, USA, Aug. 2014), 701–710.

22. Zhu, Z., Xu, S., Qu, M. & Tang, J. GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding. *The World Wide Web Conference on - WWW '19*. arXiv: 1903.00757, 2494–2504 (2019).

23. Bawa, M., Condie, T. & Ganesan, P. *LSH forest: self-tuning indexes for similarity search* in *Proceedings of the 14th international conference on World Wide Web - WWW '05* (ACM Press, Chiba, Japan, 2005), 651.

24. Koren, S., Walenz, B. P., Berlin, K., *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27,** 722–736 (May 2017).

25. Ono, Y., Asai, K. & Hamada, M. PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics* **29,** 119–121 (Jan. 2013).

26. Stöcker, B. K., Köster, J. & Rahmann, S. SimLoRD: Simulation of Long Read Data. *Bioinformatics (Oxford, England)* **32,** 2704–2706 (Sept. 2016).

27. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics (Oxford, England)* **30,** 3399–3401 (Dec. 2014).

28. Watson, M., Thomson, M., Risse, J., *et al.* PoRe: An R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics (Oxford, England)* **31** (Aug. 2014).

29. Leggett, R. M., Heavens, D., Caccamo, M., Clark, M. D. & Davey, R. P. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics* **32,** 142–144 (Jan. 2016).

30. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics (Oxford, England)* **34,** 2666–2669 (Aug. 2018).

31. Koren, S., Schatz, M. C., Walenz, B. P., *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* **30.** Number: 7 Publisher: Nature Publishing Group, 693–700 (July 2012).

32. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics (Oxford, England)* **30,** 3004–3011 (Nov. 2014).

33. Goodwin, S., Gurtowski, J., Ethe-Sayers, S., *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research* **25,** 1750–1756 (Nov. 2015).

34. Boža, V., Brejová, B. & Vinař, T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE* **12.** Publisher: Public Library of Science, e0178751 (May 2017).

35. David, M., Dursi, L. J., Yao, D., Boutros, P. C. & Simpson, J. T. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics (Oxford, England)* **33,** 49–55 (Jan. 2017).

36. Teng, H., Cao, M. D., Hall, M. B., *et al.* Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience* **7** (May 2018).

37. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20,** 129 (June 2019).

38. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13,** 238 (Sept. 2012).

39. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32,** 2103–2110 (July 2016).

40. Sović, I., Šikić, M., Wilm, A., *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications* **7,** 1–11 (Apr. 2016).

41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34,** 3094–3100 (Sept. 2018).

42. Maric, J., Sovic, I., Krizanovic, K., Nagarajan, N. & Sikic, M. *Graphmap2 - splice-aware RNA-seq mapper for long reads* preprint (Bioinformatics, July 2019).

43. Sedlazeck, F. J., Rescheneder, P., Smolka, M., *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* **15,** 461 (June 2018).

44. English, A. C., Richards, S., Han, Y., *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLOS ONE* **7.** Publisher: Public Library of Science, e47768 (Nov. 2012).

45. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15,** 211 (June 2014).

46. Bashir, A., Klammer, A., Robins, W. P., *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology* **30,** 701–707 (July 2012).

47. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics (Oxford, England)* **32,** 1009–1015 (Apr. 2016).

48. Faino, L., Seidl, M. F., Datema, E., *et al.* Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome. *mBio* **6.** Publisher: American Society for Microbiology Section: Research Article (Sept. 2015).

49. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12.** Number: 8 Publisher: Nature Publishing Group, 733–735 (Aug. 2015).

50. Clavijo, B. J., Venturini, L., Schudoma, C., *et al.* An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research* **27,** 885–896 (May 2017).

51. Zimin, A. V., Puiu, D., Luo, M.-C., *et al.* Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* **27,** 787–792 (May 2017).

52. Jiao, Y., Peluso, P., Shi, J., *et al.* Improved maize reference genome with single-molecule technologies. *Nature* **546.** Number: 7659 Publisher: Nature Publishing Group, 524–527 (June 2017).

53. Wei, C., Yang, H., Wang, S., *et al.* Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. *Proceedings of the National Academy of Sciences* **115.** Publisher: National Academy of Sciences Section: PNAS Plus, E4151–E4158 (May 2018).

54. Zhang, L., Hu, J., Han, X., *et al.* A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 1494 (Apr. 2019).

55. Belser, C., Istace, B., Denis, E., *et al.* Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* **4.** Number: 11 Publisher: Nature Publishing Group, 879–887 (Nov. 2018).

56. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology* **13,** e1005595 (June 2017).

57. Chin, C.-S., Peluso, P., Sedlazeck, F. J., *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13.** Number: 12 Publisher: Nature Publishing Group, 1050–1054 (Dec. 2016).

58. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* **27,** 737–746 (May 2017).

59. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37.** Number: 5 Publisher: Nature Publishing Group, 540–546 (May 2019).

60. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* **17.** Number: 2 Publisher: Nature Publishing Group, 155–158 (Feb. 2020).

61. Jain, M., Olsen, H. E., Turner, D. J., *et al.* Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology* (Mar. 2018).

62. Jain, M., Koren, S., Miga, K. H., *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36,** 338–345 (2018).

63. Stancu, M. C., Roosmalen, M. J., Renkens, I., *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications* **8,** 1326 (Nov. 2017).

64. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35.** Publisher: Oxford Academic, 2907–2915 (Sept. 2019).

65. Coster, W. D., Rijk, P. D., Roeck, A. D., *et al.* Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, gr.244939.118 (June 2019).

66. Cereb, N., Kim, H. R., Ryu, J. & Yang, S. Y. Advances in DNA sequencing technologies for high resolution HLA typing. *Human Immunology* **76,** 923–927 (Dec. 2015).

67. Ammar, R., Paton, T. A., Torti, D., Shlien, A. & Bader, G. D. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research* **4** (May 2015).

68. Albrecht, V., Zweiniger, C., Surendranath, V., *et al.* Dual redundant sequencing strategy: Full-length gene characterisation of 1056 novel and confirmatory HLA alleles. *HLA* **90,** 79–87 (Aug. 2017).

69. Wang, M., Beck, C. R., English, A. C., *et al.* PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC genomics* **16,** 214 (Mar. 2015).

70. Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R. & Timp, W. Nanopore sequencing detects structural variants in cancer. *Cancer Biology & Therapy* **17.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15384047.2016.1139236, 246–253 (Mar. 2016).

71. Tang, H., Kirkness, E. F., Lippert, C., *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *American Journal of Human Genetics* **101,** 700–715 (Nov. 2017).

72. Castro-Wallace, S. L., Chiu, C. Y., John, K. K., *et al.* Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Scientific Reports* **7.** Number: 1 Publisher: Nature Publishing Group, 18022 (Dec. 2017).

73. Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y. & Tighe, S. W. Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer. *Journal of Biomolecular Techniques : JBT* **28,** 2–7 (Apr. 2017).

74. Quick, J., Loman, N. J., Duraffour, S., *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530.** Number: 7589 Publisher: Nature Publishing Group, 228–232 (Feb. 2016).

75. Faria, N. R., Sabino, E. C., Nunes, M. R. T., *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine* **8** (Sept. 2016).

76. Kafetzopoulou, L. E., Pullan, S. T., Lemey, P., *et al.* Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science* **363.** Publisher: American Association for the Advancement of Science Section: Report, 74–77 (Jan. 2019).

77. Septimus, E. J. & Owens, R. C. Need and potential of antimicrobial stewardship in community hospitals. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **53 Suppl 1,** S8–S14 (Aug. 2011).

78. Ashton, P. M., Nair, S., Dallman, T., *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology* **33.** Number: 3 Publisher: Nature Publishing Group, 296–300 (Mar. 2015).

79. Bradley, P., Gordon, N. C., Walker, T. M., *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications* **6.** Number: 1 Publisher: Nature Publishing Group, 10063 (Dec. 2015).

80. Schmidt, K., Mwaigwisya, S., Crossman, L. C., *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *The Journal of Antimicrobial Chemotherapy* **72,** 104–114 (Jan. 2017).

81. Lemon, J. K., Khil, P. P., Frank, K. M. & Dekker, J. P. Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates. *Journal of Clinical Microbiology* **55,** 3530–3543 (Dec. 2017).

82. Greninger, A. L., Naccache, S. N., Federman, S., *et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine* **7,** 99 (Sept. 2015).

83. Charalampous, T., Kay, G. L., Richardson, H., *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology* **37.** Number: 7 Publisher: Nature Publishing Group, 783–792 (July 2019).

84. Votintseva, A. A., Bradley, P., Pankhurst, L., *et al.* Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *Journal of Clinical Microbiology* **55.** Publisher: American Society for Microbiology Journals Section: Mycobacteriology and Aerobic Actinomycetes, 1285–1298 (May 2017).

85. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics,* **3.** Publisher: Microbiology Society, e000132 (2017).

86. Koren, S., Harhay, G. P., Smith, T. P., *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* **14,** R101 (Sept. 2013).

87. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology* **38.** Number: 6 Publisher: Nature Publishing Group, 701–707 (June 2020).

88. Fichot, E. B. & Norman, R. S. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* **1,** 10 (Mar. 2013).

89. Mosher, J. J., Bowman, B., Bernberg, E. L., *et al.* Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *Journal of Microbiological Methods* **104,** 59–60 (Sept. 2014).

90. Benítez-Páez, A., Portune, K. J. & Sanz, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience* **5,** 4 (2016).

91. Weirather, J. L., de Cesare, M., Wang, Y., *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6,** 100 (2017).

92. Zhang, G., Sun, M., Wang, J., *et al.* PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. *The Plant Journal: For Cell and Molecular Biology* **97,** 296–305 (Jan. 2019).

93. Soneson, C., Yao, Y., Bratus-Neuenschwander, A., *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 3359 (July 2019).

94. Tombácz, D., Csabai, Z., Oláh, P., *et al.* Characterization of Novel Transcripts in Pseudorabies Virus. *Viruses* **7,** 2727–44 (May 2015).

95. Deininger, P., Morales, M. E., White, T. B., *et al.* A comprehensive approach to expression of L1 loci. *Nucleic Acids Research* **45,** e31 (Mar. 2017).

96. Workman, R. E., Tang, A. D., Tang, P. S., *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods* **16,** 1297–1305 (Dec. 2019).

97. Liu, H., Begik, O., Lucas, M. C., *et al.* Accurate detection of m 6 A RNA modifications in native RNA sequences. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 4079 (Sept. 2019).

98. Parker, M. T., Knop, K., Sherwood, A. V., *et al.* Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife* **9** (eds Wan, Y. & Hardtke, C. S.) Publisher: eLife Sciences Publications, Ltd, e49658 (Jan. 2020).

99. Flusberg, B. A., Webster, D. R., Lee, J. H., *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7.** Number: 6 Publisher: Nature Publishing Group, 461–465 (June 2010).

100. Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences* **110.** Publisher: National Academy of Sciences Section: Biological Sciences, 18904–18909 (Nov. 2013).

101. Schreiber, J., Wescoe, Z. L., Abu-Shumays, R., *et al.* Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences* **110.** Publisher: National Academy of Sciences Section: Biological Sciences, 18910–18915 (Nov. 2013).

102. Simpson, J. T., Workman, R. E., Zuzarte, P. C., *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14,** 407–410 (Feb. 2017).

103. Rand, A. C., Jain, M., Eizenga, J. M., *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods* **14,** 411–413 (Feb. 2017).

104. Liu, Q., Fang, L., Yu, G., *et al.* Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 2449 (June 2019).

105. Ni, P., Huang, N., Luo, F. & Wang, J. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning (Aug. 2018).

106. Yuen, Z. W.-S., Srivastava, A., Daniel, R., *et al.* Systematic benchmarking of tools for CpG methylation detection from Nanopore sequencing. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.10.14.340315 (Nov. 2020).

107. Euskirchen, P., Bielle, F., Labreche, K., *et al.* Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica* **134,** 691–703 (Nov. 2017).

108. Lee, I., Razaghi, R., Gilpatrick, T., *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods* **17.** Number: 12 Publisher: Nature Publishing Group, 1191–1199 (Dec. 2020).

109. Liu, Y., Cheng, J., Siejka-Zielińska, P., *et al.* Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biology* **21,** 54 (Mar. 2020).

110. Jain, M., Fiddes, I. T., Miga, K. H., *et al.* Improved data analysis for the MinION nanopore sequencer. *Nature Methods* **12.** Number: 4 Publisher: Nature Publishing Group, 351–356 (Apr. 2015).

111. Loman, N. J. & Watson, M. Successful test launch for nanopore sequencing. *Nature Methods* **12,** 303–304 (Apr. 2015).

112. Magi, A., Semeraro, R., Mingrino, A., Giusti, B. & D'Aurizio, R. Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in Bioinformatics* **19,** 1256–1272 (2018).

113. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (Oct. 2012).

114. Li, H., Handsaker, B., Wysoker, A., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (Aug. 2009).

115. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (Mar. 2010).

116. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26,** 2204–2207 (Sept. 2010).

117. Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLOS Computational Biology* **14,** e1006583 (Nov. 2018).

118. ONT. *R10.3: the newest nanopore for high accuracy nanopore sequencing – now available in store* Section: News. 2020.

119. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nature Methods* **13,** 751–754 (July 2016).

120. Stoiber, M. H., Quick, J., Egan, R., *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv,* 094672 (Apr. 2017).

121. Ferguson, J. M. & Smith, M. A. SquiggleKit: a toolkit for manipulating nanopore signal data. *Bioinformatics* **35.** Publisher: Oxford Academic, 5372–5373 (Dec. 2019).

122. Gamaarachchi, H., Lam, C. W., Jayatilaka, G., *et al.* GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinformatics* **21,** 343 (Aug. 2020).

123. Rohrandt, C., Kraft, N., Gießelmann, P., *et al. Nanopore SimulatION – a raw data simulator for Nanopore Sequencing* in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Dec. 2018), 1–8.

124. Li, Y., Wang, S., Bi, C., *et al.* DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing. *Bioinformatics* **36.** Publisher: Oxford Academic, 2578–2580 (Apr. 2020).

125. Gonzalez, R. C. & Woods, R. E. *Digital Image Processing (3rd Edition)* (Prentice-Hall, Inc., USA, 2006).

126. Reinert, K., Dadi, T. H., Ehrhardt, M., *et al.* The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. *Journal of Biotechnology* **261,** 157–168 (Nov. 2017).

127. Lin, J., Keogh, E., Lonardi, S. & Chiu, B. *A symbolic representation of time series, with implications for streaming algorithms* in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* (Association for Computing Machinery, New York, NY, USA, June 2003), 2–11.

128. Šošić, M. & Šikić, M. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* **33.** Publisher: Oxford Academic, 1394–1395 (May 2017).

129. Schreiber, J. & Karplus, K. Analysis of nanopore data using hidden Markov models. *Bioinformatics* **31,** 1897–1903 (June 2015).

130. Marwal, A., Sahu, A. K. & Gaur, R. K. in *Animal Biotechnology* (eds Verma, A. S. & Singh, A.) 289–305 (Academic Press, San Diego, Jan. 2014).

131. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Research* **26,** 115–138 (Sept. 2018).

132. Gatchel, J. R. & Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews. Genetics* **6,** 743–755 (Oct. 2005).

133. Paulson, H. Repeat expansion diseases. *Handbook of Clinical Neurology* **147,** 105–123 (2018).

134. Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65,** 905–914 (May 1991).

135. Van Blitterswijk, M., DeJesus-Hernandez, M., Niemantsverdriet, E., *et al.* Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. *The Lancet. Neurology* **12,** 978–988 (Oct. 2013).

136. Xi, Z., Zinman, L., Moreno, D., *et al.* Hypermethylation of the CpG island near the G4C2 repeat in ALS with a C9orf72 expansion. *American Journal of Human Genetics* **92,** 981–989 (June 2013).

137. Russ, J., Liu, E. Y., Wu, K., *et al.* Hypermethylation of repeat expanded C9orf72 is a clinical and molecular disease modifier. *Acta Neuropathologica* **129,** 39–52 (Jan. 2015).

138. Mizielinska, S., Grönke, S., Niccoli, T., *et al.* C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. *Science (New York, N.Y.)* **345,** 1192–1194 (Sept. 2014).

139. Dashnow, H., Lek, M., Phipson, B., *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biology* **19,** 121 (2018).

140. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Medicine* **9,** 65 (2017).

141. O'Rourke, J. G., Bogdanik, L., Muhammad, A. K. M. G., *et al.* C9orf72 BAC Transgenic Mice Display Typical Pathologic Features of ALS/FTD. *Neuron* **88,** 892–901 (Dec. 2015).

142. Hornstra, I. K., Nelson, D. L., Warren, S. T. & Yang, T. P. High resolution methylation analysis of the FMR1 gene trinucleotide repeat region in fragile X syndrome. *Human Molecular Genetics* **2,** 1659–1665 (Oct. 1993).

143. Xi, Z., Zhang, M., Bruni, A. C., *et al.* The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathologica* **129,** 715–727 (May 2015).

144. Lyons, J. I., Kerr, G. R. & Mueller, P. W. Fragile X Syndrome: Scientific Background and Screening Technologies. *The Journal of molecular diagnostics: JMD* **17,** 463–471 (Sept. 2015).

List of Figures

List of Tables

List of Listings

Declaration of Authorship

I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure. I agree to have my thesis examined by a plagiarism examination software.

Berlin, 01.12.2021

Pay Giesselmann

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses a customized version of the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.