# Genomics applications of nanopore long-read sequencing for small to large sized genomes

Liem, M.

# Genomics applications of nanopore long-read sequencing for small to large sized genomes

# Genomics applications of nanopore long-read sequencing for small to large sized genomes

Cover art: Elephant by Michael Liem.

Referring to the largest sequencing reads among sequencing technologies.
The plethora of animal illustrations decorating the elephant represents the Oxford Nanopore Technologies slogan “*To enable the analysis of anything, by anyone, anywhere.*”

| | |
|---|---|
| Lay-out: | Ontwerpstudio Carlijn Hoogenboom |
| Printed by: | Gildeprint - Enschede |
| ISBN: | 978-94-6496-089-1 |

# Genomics applications of nanopore long-read sequencing for small to large sized genomes

## Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 17 april 2024
klokke 11:15

door

Michael Liem
Geboren te Alphen aan den Rijn, Nederland
in 1987

|  |  |  |
|---|---|---|
| Promotors: | Prof. dr. H.P. Spaink | Leiden University |
|  | Prof. dr. ir. F.J. Verbeek | Leiden University |
| Co-promotor: | Dr. C.V. Henkel | Norwegian University of Life Sciences |
| Promotiecommissie: | Prof. dr. A.H. Meijer | Leiden University |
|  | Prof. dr. V. van Noord | Leiden University |
|  | Dr. K.J. Wolstencroft | Leiden University |
|  | Dr. J. Masschelein | Katholieke Universiteit Leuven |
|  | Prof. dr. ir. D. de Ridder | Wageningen University & Research |

“You must be shapeless, formless, like water.
When you pour water in a cup, it becomes the cup.
When you pour water in a bottle, it becomes the bottle.
When you pour water in a teapot, it becomes the teapot.
Water can drip and it can crash.
Become like water my friend.”

– Bruce Lee

# — Contents

— Chapter 1

# Introduction and thesis outline

# Applications of DNA sequencing

## Genome sequencing

Nucleotide sequencing has revolutionized the discovery of genomic content and has enabled the scientific community to unveil the genetic code for a large range of organisms. Sequencing started in 1965 when the full sequence and structure of the first tRNA was detected, a molecule of 77 nucleotides[1]. Innovations in the next decade enabled completion of viral genomes in the kilobasepair (Kbp) range and at the end of the 1990's this was already extended to megabasepair (Mbp) size genomes[2]. Sequencing technologies provide increasing data volumes because of advancements in the speed at which nucleotide sequences are detected, the effectiveness of library chemicals, and the degree of parallelization.

Incredible efforts have been made to collect those datasets, as well as perform downstream analyses such as assemblies, annotations, and variant identification. These in turn enabled many biological applications and allowed the scientific community to investigate areas that had remained unreachable. An example of such an area is assembly; in the application known as 'genome assembly' whole genomes are reconstructed based upon the overlap of small fragments. Those fragments, known as 'reads', are the readout of fragmented DNA molecules. Homology between reads allows resolving fragmentation into functional units such as chromosomes or plasmids. To accumulate adequate evidence for the reconstruction of those fragmented datasets, adequate sequencing 'depth' is required, which has pushed data generation to its current magnitude. Around 2008 whole genome (WGS) and whole exome sequencing (WES) have taken a giant leap and became a dominant factor in generating large datasets among multiple biomedical data science disciplines (Figure 1) [3].

Unprecedented increase of WGS and WES data generation compared to other data science disciplines.

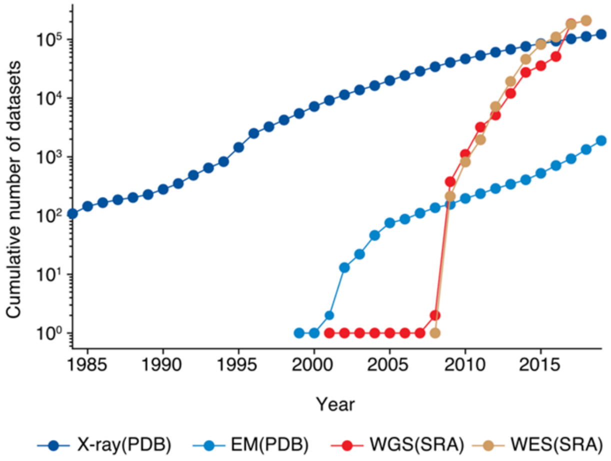

**Figure 1 Cumulative number of datasets being generated for whole genome sequencing (WGS – SRA) and whole exome sequencing (WES – SRA).** In comparison to molecular structure datasets such as X-ray and electron microscopy (EM – PDB), generating sequencing data has become the dominant player[3].

Currently we can generate terabase datasets in just a couple of days. Large amounts of data and increased computational capacity have allowed us to extend our reach in analyzing genomes that have previously been unresolvable due to their size and structural complexity.

## Functional genomics

Unraveling genomic content has been the initial motivation for developing sequencing technologies and bioinformatics. Furthermore, understanding the transcriptome is essential to investigate functional units of the genome and the molecular interaction that regulate cells and tissues. It is key to collect all transcript isoforms, structures such as splice variants and posttranscriptional modification, and lastly the change of expression of those transcripts during cellular and tissue functioning[4]. However, conserved genetic information in the DNA molecules combined with the functional effects through expression of transcripts and the translation to proteins is insufficient to explain the entire process of how cells and tissues derive their phenotype.

Currently, the study of elements binding to DNA and DNA modifications (e.g. epigenetics) provides insight on the use and function of the genomic architecture that was not foreseen when the first full genome sequencing projects provided a detailed view of the nucleotide level. Among others, ENCODE (Encyclopedia of DNA Elements) has put in enormous effort into annotating the human genome and revealed that most functional annotations have a regulative nature and are not protein-coding[5]. This adds to the notion that the mechanism of regulation is just as important compared to the unraveling of the structural genomic code or cataloguing transcript variation among cells and tissues.

## Diagnostics

A current gold-standard method is targeted sequencing, where we amplify a region of interest and sequence the amplified product. These efforts are used for, among others, variant detection, determination of structural variation, identification of isoforms and full-length mRNA sequencing. These sequencing data provide a detailed targeted insight on a region of interest and have proved to be a powerful tool for diagnostics. For this setup relatively small datasets (short regions and high coverage) are required to determine the genomic content. However, amplification introduces bias towards shorter molecules since shorter molecules are amplified faster compared to longer molecules. Hence the final amplified sample contains a bias towards the number of molecules in favor of shorter lengths, hindering quantitative analyses. Therefore, amplification-free sequencing does not only simplify library preparation protocols, in addition it circumvents skewness and can generate unbiased libraries facilitating the quantification of sequencing samples.

Furthermore, disease-related genomic aberrations are not solely bound to single target variations, therefore a more comprehensive assessment of the genome is often required. WGS provides information on the complete genome, including coding and non-coding regions, and offers a better approach towards copy number variation, genomic rearrangements, and other structural variations, causing WGS data to have a more predictive nature. Additionally, on a general note simplifying preparations decreases library preparation complexity in terms of time and machinery. This in turn allows in-field sequencing and bridges a gap between sample collection and identification for areas without the availability of high-tech equipment.

### Metagenomics

Metagenomics is the readout of a set of sequencing reads from a pool of input DNA, with the aim of reconstructing which species are present within a sample. Furthermore, not all organisms in a complex bacterial sample can be cultured under laboratory conditions, therefore sequencing is needed to detect their presence and abundance. Because of the inequal quantities of represented organisms in a metagenomics sample, it is challenging to assemble all their complete genomes. However, with highly improved sequencing technologies, described below, assembly of genomes of poorly represented unculturable organisms using metagenomics comes within reach.

# — Brief summary of sequencing techniques

## Sanger

Sanger sequencing relies on amplification of the input DNA molecule, denaturating double stranded DNA molecules to separate the strands, and the subsequent annealing of primers to single strand DNA molecules that form the start site of the sequencing (Figure 2 – step 1). Primer annealed single strand molecules are exposed to dNTPs (deoxynucleotide triphosphate) mixed with fluorescently labeled ddNTPs (dideoxy nucleotides) at low concentration. ddNTPs have both hydroxyl groups absent from the sugar backbone molecule to terminate the reaction when reached by the polymerase. This results in amplified DNA molecules of different lengths that require gel electrophoresis to determine the final incorporated nucleotide, and thereby the full sequence (Figure 2 – step 2 and 3). This imposes a major limitation towards sequencing many, or longer molecules. Furthermore, sequencing techniques based on amplification do not allow direct readout of the input molecule, making them susceptible to bias introductions.
Finally, classical Sanger sequencing is designed to readout small DNA molecules, and the technique is unable to scale towards molecules of their original size, ranging from millions (for small bacterial genomes) to billions of base pairs (for human genomes or larger).

Three-step schematic overview of Sanger sequencing

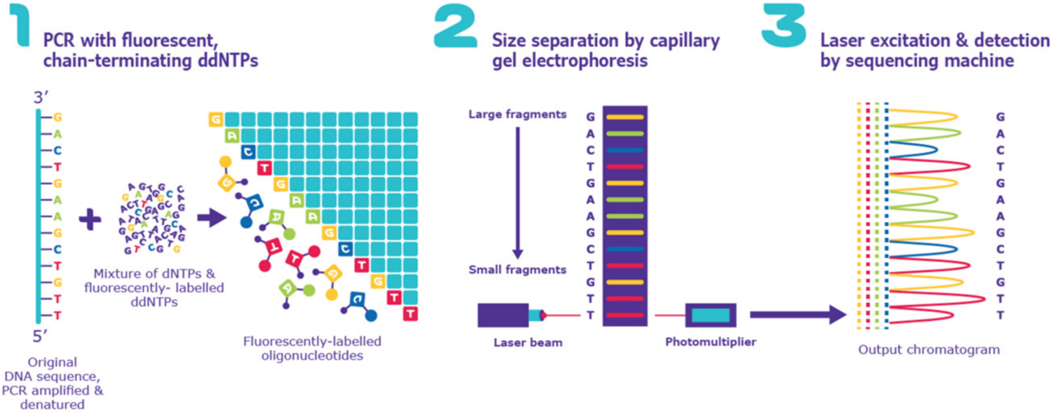
1
PCR with fluorescent, chain-terminating ddNTPs
3′
G
A
C
T
G
A
A
G
C
T
G
T
T
5′
Original DNA sequence, PCR amplified & denatured
Mixture of dNTPs & fluorescently- labelled ddNTPs
Fluorescently-labelled oligonucleotides
2
Size separation by capillary gel electrophoresis
Large fragments
Small fragments
G
A
C
T
G
A
A
G
C
T
G
T
T
Laser beam
Photomultiplier
3
Laser excitation & detection by sequencing machine
G
A
C
T
G
A
A
G
C
T
G
T
T
Output chromatogram

Figure 2 **Overview of the Sanger sequencing technique. 1**) Denatured DNA molecule, primers form sequencing start site. DNA fragments are fluorescently labeled. **2**) Size separation using gel electrophoresis and incorporated nucleotide identification. **3**) fluorescently labeled nucleotides are excited by laser allowing readout of the sequence. Figure adopted from Sigmaaldrich – Sanger sequencing steps and method[9].

## Illumina - Next generation sequencing (NGS)

The name ‘next generation’ highlights the ability to overcome small quantity DNA sequencing. Compared to classical Sanger sequencing, NGS allows sequencing of millions of small molecules (max. 300 bp) in parallel. In short, double stranded input DNA is fragmented, denatured to single strand DNA and two distinct adapters (oligonucleotides) are ligated to either side of the small fragments. On a glass plate, also known as the flow cell chip, sequences that are complementary to the two distinct adapters are attached to the flow cell surface, allowing input DNA adapters to bind (Figure 2 – step 1). Then single stranded DNA is amplified using bridge amplification (Figure 3 – step 2, 3, 4 and 5), yielding local clusters of thousands of identical molecules, ready for sequencing by synthesis (Figure 3 – step 6). Fluorescently labeled nucleotides bind to their complementary part on the strand by DNA polymerase, and using a laser the last incorporated fluorescent label is excited, which serves as the sequencing signal. This generates digital images (Figure 3 – final illustration) and through image analysis the final sequence is determined.

This sequencing technique is highly accurate, less then 0,1% of all sequenced nucleotides is classified incorrectly. This is mainly due to cluster generation, which boosts the sequencing signal significantly. However, this amplification also limits the length of DNA fragments that can be analyzed. The major drawback for data generated by this technique is therefore the millions of small fragments that must be aligned to increase the sequence length to resolve the original input DNA sequence. Such alignment tasks are resource-heavy and require complicated algorithms to overcome intricate genome structures. Furthermore, genomic regions such as repeats or low complexity regions are impossible to reconstruct using next generation sequencing data altogether, since those short reads do not provide sufficient overlap to elongate sequences that contain repetitiveness longer than the sequencing read itself.

Illumina sequencing technique converts incorporated fluoresently labeld bases in to digital images

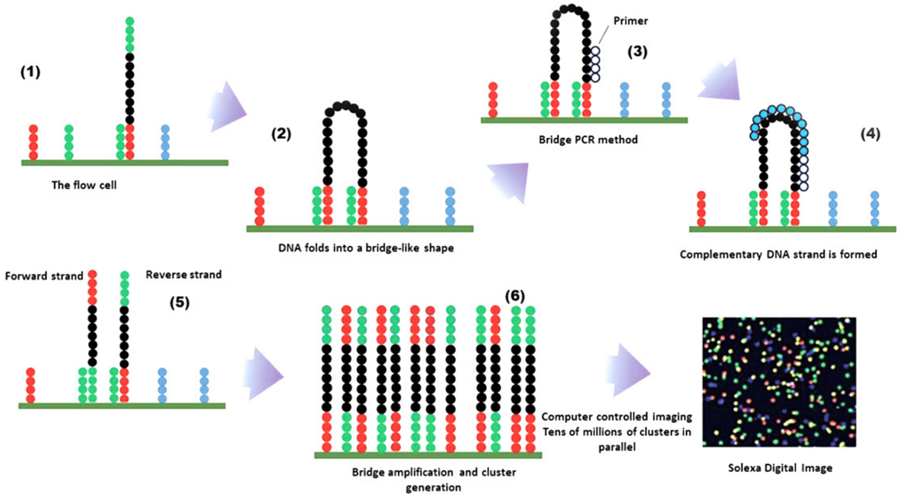
(1)
The flow cell
(2)
DNA folds into a bridge-like shape
Primer
(3)
Bridge PCR method
(4)
Complementary DNA strand is formed
Forward strand
Reverse strand
(5)
(6)
Bridge amplification and cluster generation
Computer controlled imaging Tens of millions of clusters in parallel
Solexa Digital Image

Figure 3 **Schematic overview of Illumina sequencing technology.** 1) input DNA is attached to the flow cell chip, 2, 3, 4, 5 and 6) bridge amplification generates clusters to increase the sequencing signal. **Final illustration**) sequencing signal is captured through digital imaging using the excitation of fluorescently labeled nucleotides, image analysis results in the final sequence. Figure adopted from[6].

## Pacific Biosciences (PacBio) sequencing

High throughput sequencing techniques that can overcome short read-lengths are referred to as third-generation sequencing. In contrast to Illumina sequencing, they operate on single DNA molecules. Since this technique captures signals originating from a single molecule the sequencing signal is much weaker compared to Illumina sequencing, hence more errors are introduced to the final readout. One of the most prominent platforms is PacBio single molecule real-time (SMRT) sequencing. SMRT libraries are generated through ligating adapters circularizing the input DNA and allowing a primer to bind to one of the adapters as the start site for a polymerase. The library is mobilized to millions of tiny wells, also known as zero mode wave guides (ZMW's) (Figure 4 A), such that every ZMW contains a single molecule. A single polymerase is attached at the bottom of the well and incorporates fluorescently labeled nucleotides of which the emitted light is detected using lasers and is measured in real-time (Figure 4 B). Millions of ZMW's are measured in parallel allowing the production of Gbp datasets and generating Kbp read-lengths during a single run.

Due to the circularized nature of this technique PacBio sequencing can generate two kinds of reads. The first kind is circular consensus sequences (currently marketed as HiFi reads), this technique can read a single molecule multiple times using the sequence multitude to correct for randomly introduced errors. Although high accuracy reads find their strength in circularization of single molecules, it simultaneously limits the maximum read-length and introduces an upper limit. The second kind is called continuous long read sequencing and can reach maximum read-lengths (up to 70 Kbp in length) at the cost of sequence accuracy. Since only a single read-out is generated during this mode no consensus can be derived.

PacBio readout of single molecules from zero mode wave guides generates super long reads at the expense of accuracy

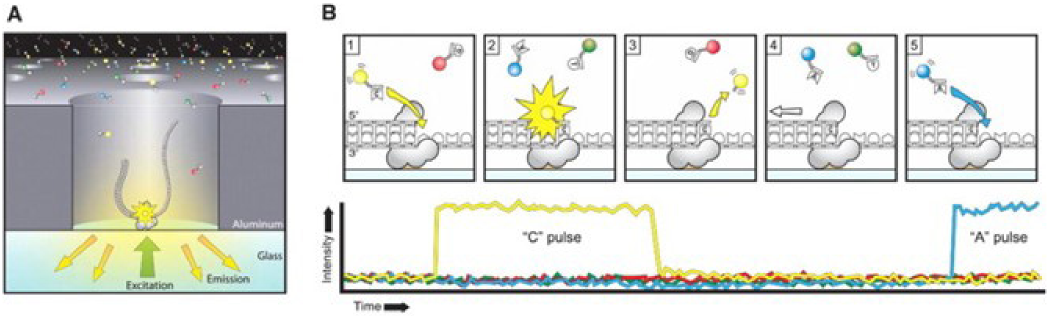

Figure 4 A) illustration of a zero mode wave guide, a small well with a polymerase attached to the bottom. B) fluorescently labeled nucleotides are incorporated by the polymerase, incorporated nucleotides emit a light signal that is captured by a laser system, converting intro a graph tracing the signal intensity over time, allowing the readout of the input DNA. Figure adopted from[7].

## Nanopore sequencing

A relative newcomer to the third-generation sequencing field is Oxford Nanopore Technologies (ONT). Their technique utilizes a distinct method compared to the previously described techniques, which all rely on DNA polymerase. DNA molecules are guided through a pore protein embedded into a membrane. An electrical current is applied across the membrane (Figure 5 A). ONT uses the profile of electrical current distortions (squiggle profiles, Figure 5 B) to differentiate nucleotides as they pass through the pore protein. DNA molecules are prepared by ligating adapter sequences to both ends of the DNA fragment, motor proteins (Figure 5 A, motor protein in yellow) are attached to the molecule and control the speed at which the readout is performed. Since the determination of the final sequencing read is based on algorithmic interpretation of the squiggle profiles, controlling the speed is a crucial step. When nucleotides move rapidly through the pore the algorithm might not be able to differentiate nucleotides within the squiggle profile leading to error introduction. Currently, around 450 nucleotides per second move through the pore in a sliding window setup, which means multiple nucleotides occupy the pore protein at the same time. Hence signals captured in the squiggle profiles are electrical current distortions of a nucleotide 5-mer[8] (Figure 5 A, 5-mer situated at the narrowest region of the pore), therefore leading to a multiplicity of states every time a base enters or leaves the pore shaft, at the top or bottom side, respectively. Five bases per moment in time and a four-base model yields $4^5$=1024 different states, and the addition of base modifications only increases the total number of states exponentially, making basecalling a challenging machine learning problem.

Pulling a single DNA molecule through the nanopore distorts the ion current, interpretation of those distortion signals yields the final readout

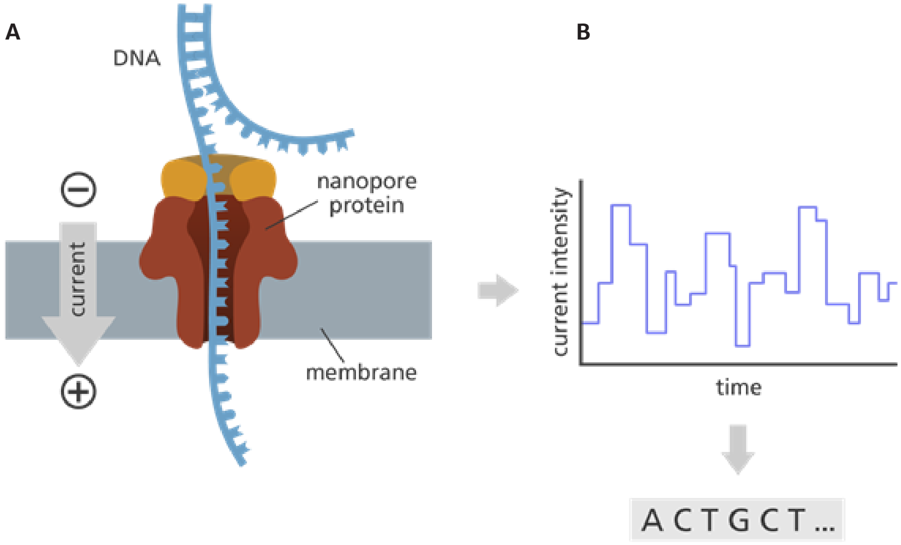

Figure 5 **Schematic overview of Oxford Nanopore Technologies.** A) nanopore embedded into a membrane with an ionic current through the nanopore, a motor protein (in yellow) controls the speeds at which DNA molecules are pulled through the pore to facilitate sequence accuracy. B) ionic current distortions are converted to graphs, called squiggle plots, which form the input data for machine learning algorithms to basecall the final sequencing readout. Figure adopted from[9].

## Long read quality

The major benefit of third-generation sequencing techniques is the length of sequencing reads. With a theoretical indefinite upper limit (for ONT) this technique is essential for unraveling the structure of large genomes that have been unreachable up until this point. The downside of long read sequencing techniques (both ONT and PacBio) is that they have struggled to reach read accuracy comparable to gold-standard sequencing techniques such as Illumina. The quality of reads is expressed using Phred scores, which is a logarithmic metric to indicate the number of miscalled nucleotides. Q10 stands for 10% misclassified bases in a sequencing read, Q20 for 1% error and so on. Current ONT sequencing flow cells and chemistry yield read-lengths around 10-100 Kbp (maximum reported reads range between 2-4 Mbp[28]) and deliver data in excess of around Q20. ONT originally delivered data with over 30% misclassifications (around 2015), indicating that within a relatively short time-span the quality of this sequencing technique has dramatically improved.

## Assembly evolution and genome complexity

Assembly has been an evolving strategy in reconstructing genomes from a set of smaller reads. There have been several approaches that have facilitated the scientific community aiming to resolve the genetic makeup from sequencing data. In the beginning hierarchical approaches were used, this strategy typically uses large insert BAC clones, where numerous clones cover the insert sequence in a tiling-path manner. Then a minimal tiling-path represents the consensus sequence of the insert sequence, hereafter, insert sequences are manually closed yielding the final assembly. This top-down approach utilizes preexisting knowledge that guides the assembly task at hand.

Due to high-throughput parallel shot-gun sequencing and increased computational capabilities bottom-up approaches became within reach. Here no preexisting knowledge is required for the task at hand; piece back millions of reads to reconstruct a genome. First, the naïve approach. Theoretically one could find overlaps between two reads and then elongate those into larger contigs (contiguous sequences) until the full genome is assembled. Using a four letter alphabet {A,T,C and G} and 20 bases overlap, the uniqueness with which such a 20-mer is found in the genome equals to $4^{20}$, which translates to once every 1 x $10^{12}$ base pairs. However, this uniqueness only upholds for exact matches and randomly distributed nucleotides. Since the distribution of nucleotides in biological sequences are not random, 20-mer overlaps are found much more frequently. The data quantities generated by NGS sequencing techniques and the erroneous nature of long read sequencing data cause assemblies to reach complexity levels to a point that simple base comparison becomes computationally infeasible.

Graph-based assembly approaches are a natural expansion of the naïve approach. The benefit for graph represented assemblies is to find a path that represents the genome, for which many strategies and algorithms already exists. Additionally, providing long distance information (e.g. node information from nodes other than closely positioned neighbors) graph complexities are resolved relatively simple.

Overlap-graphs construct directional graphs based on overlap. Graphs are represented by nodes and directed edges, where nodes stand for overlap and edges denote the relation between the suffix overlap of one read to the prefix overlap of another. Overlaps are determined based on all reads vs. all reads alignment, where overlap lengths are recorded and following the path of the longest overlaps determine the final assembly. Similar to the naïve approach, introduction of sequencing errors complicates the assembly by introducing additional branches to the graph. Since this assembly approach finds it fundament in all reads vs. all reads alignment resolving large data sets, particularly for large genomes and long read sequencing data, becomes computationally infeasible. Hence the bottleneck for this approach is the computationally intensive all read vs. all read alignment causing the lack of scalability towards data set quantities currently generated[15, 16]. In chapter 4 we propose an alternative version adopted from this assembly strategy. The alternative strategy limits the search space of the computationally intensive all reads vs. all reads alignment, enabling analysis of large and complex genomes using overlap-graphs.

A clever workaround managing large datasets is called a De Bruijn graph, were reads are represented by k-mers of a particular size (k-mer size ranges from 31 to 127 bases, typically k-mer sizes 31, 55, 77, 99 or 127 are applied). The De Bruijn graph is a representation of uniquely found k-mers and using a k-minus-1-kmer evaluation k-mers are linked together directionally. This simplifies the assembly graph significantly compared to overlap-graphs, allowing a better scalability for large datasets, both for large quantity data sets as well as for long-read erroneous data, and is currently the go-to method for short-read NGS assembly problems. However, the down-side of this method for longer reads is the loss of context, which is the most prominent benefit of long reads.

## Long read assembly-quality and sequence accuracy

Since assembly is a major part of the downstream analysis, quality standards are required to assess the final assembly result, where we are interested in the completeness, level of fragmentation, accuracy, resource requirements, and speed of the assembly algorithm.
For known (independently measured) genome size the completeness is based on the assembly length; the number of nucleotides in the completed assembly should meet the known genome size. The genome size can also be estimated from the sequencing data for genomes that are not yet characterized to evaluate the assembly completeness[29]. In case of multi-chromosomal genomes, the level of fragmentation should ideally not exceed the number of chromosomes or plasmids. A correct number of fragments and completeness indicates that the genome is fully assembled at the chromosome level. Contig N50 metrics are useful to evaluate the level of fragmentation. N50 is a weighted median of fragment lengths and is calculated by size sorting contigs from large to small, where the N50 indicates the size of the contig that is found at 50% of the total assembly length. A complementary analysis to the forementioned technical metrics is to quantify the functional completeness of the final assembly. In essence such a heuristic quality control, for example provided by BUSCO, runs a gene prediction on the genomic data and compares the results to the expected gene content from (closely) related organisms. This yields estimates of assembly completeness and the deleterious effects of fragmentation. Reference data sets are stored in open-source data bases and include, among others, data for vertebrates, fungi, prokaryotes and plants[17].

A well-known example, highlighting the importance of quality assessment, is the human reference genome project. Around twenty years ago the first human genome assembly was released and has set the standard for genomic applications such as alignment, variant detection, functional genomics, population genetics and epigenetic analysis[30]. The current gold-standard human reference genome (GRCh38.p13) is based on a mosaic collection of around twenty people, ~70% of which originates from a single individual[31, 32]. Hence the reference genome fails to represent the genomic content of any one person. The mosaic representation results in reference biases causing decreased accuracy for variant discovery, the association of gene-disease and other genetic analysis and left ~8% of the genome unresolved[18, 19].
In 2022 the telomere-to-telomere consortium, by combining multiple sequencing techniques and using the Verkko assembler, delivered a fully resolved, fully phased diploid representation of the human genome, where 20 of the 46 chromosomes are automatically assembled from telomer to telomer at 99.9997% accuracy[20].

Another difficulty for generating high quality assemblies arises from low sequencing read accuracy and homogeneous coverage along a genome. Downstream correction procedures, where correction methods are based on coverage multiplicity, allow per base assessment which correct misclassifications using a majority vote. Therefore, covering the same base position multiple times is essential for high accuracy assembly results, and it is important to generate sufficient and evenly spread coverage to ensure every position is covered at least three times to provide sufficient evidence during the majority vote. For this specific event overall genome coverage could be a deceptive source of information. For example, a 50 Mbp dataset from a 5 Mbp genome suggests 10x coverage, however it does not guarantee more than threefold coverage on every position.

Closer to reality, GC-enriched and repetitive regions, such as rRNA regions, appear to be difficult to sequence homogeneously and often yield read abundance on those regions departing from the theoretical coverage. Additionally, due to random sampling regions remain absent from the sequencing library independent of sequencing depth. Figure 6 shows how multiple platforms struggle to evenly cover the rRNA region, which is illustrated around 20 degrees, and recognized by a small green bar depicted in the CDS band (yellow band). For this region, Illumina MiSeq, NextSeq, Hiseq and PacBio have sharp peaks indicating a read enrichment over the repetitive rRNA region, however ONT delivers a rather even overall coverage.

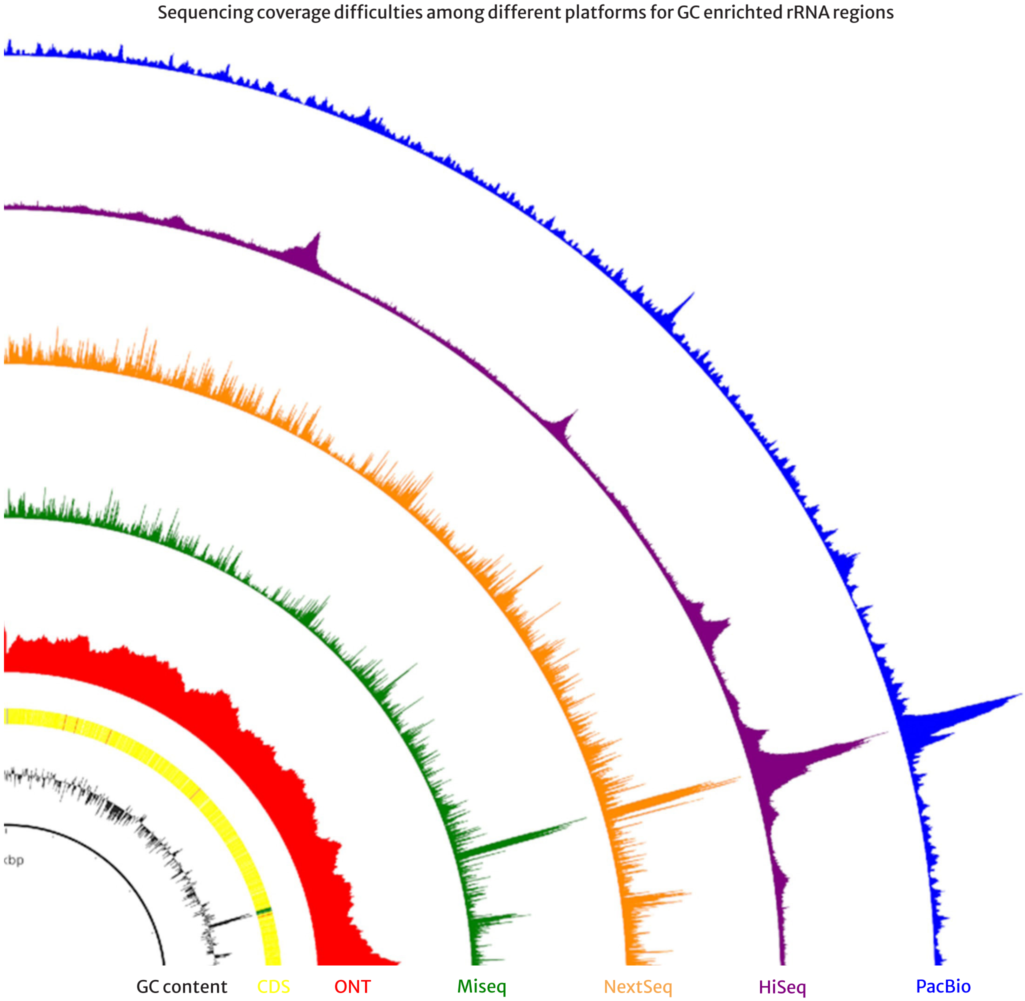

Figure 6 **Partial visualization of single chromosome assembly for Fusobacterium sp. C1.**
Circle plot bands represent, from inside to outside, GC-content (black), Coding sequence (yellow), ONT read coverage (red), Miseq coverage (green), NextSeq coverage (orange), Hiseq coverage (purple) and PacBio coverage (blue). Peak heights indicate sequence read coverage indicating that most platforms struggle to cover the genome evenly in general and particularly for the rRNA region highlighted in green inside the CDS band around 20 degrees of the circle illustration. Figure adopted from[21].

## Generating data using different platforms

The choice for a sequencing technique depends on several specifications, among others, data volume, data accuracy, read length, read count, cost, time and availability. Currently, most sequencing devices can generate high throughput data and yield incredibly large datasets. This is exemplified by the maximum output for different devices; HiSeq 4000 – 1 Tbp in 96 hours, HiSeq 2500 – 120 Gbp in 30 hours, PacBio Sequel – 60 Gbp in 30 hours, PacBio Revio – 360 Gbp in 24 hours, PacBio Onso – 150 Gbp in 48 hours, MinION – 50 Gbp in 48 hours, GridION – 250 Gbp in 72 hours, PromethION – 14 Tbp in 72 hours, see Figure 7. Maximum run-times are defined by the sequencing platforms and are usually based on depletion of biochemicals required for the sequencing run.

The maximum number of sequencing reads for Illumina sequencing is defined by the number of generated read clusters on the flow cell chip, hence a longer sequencing run yields longer sequencing reads, however the number of reads remain the same. Similar to Illumina, for PacBio sequencing the maximum number of sequencing reads is limited by the availability of zero mode wave guides on the flow cell. However, for Oxford Nanopore Technologies the maximum number of reads is restricted by the size and volume of input DNA together with the speed at which a DNA molecule is pulled through the pore. Hence for fixed data volumes the maximum number of reads is proportional to the read length.

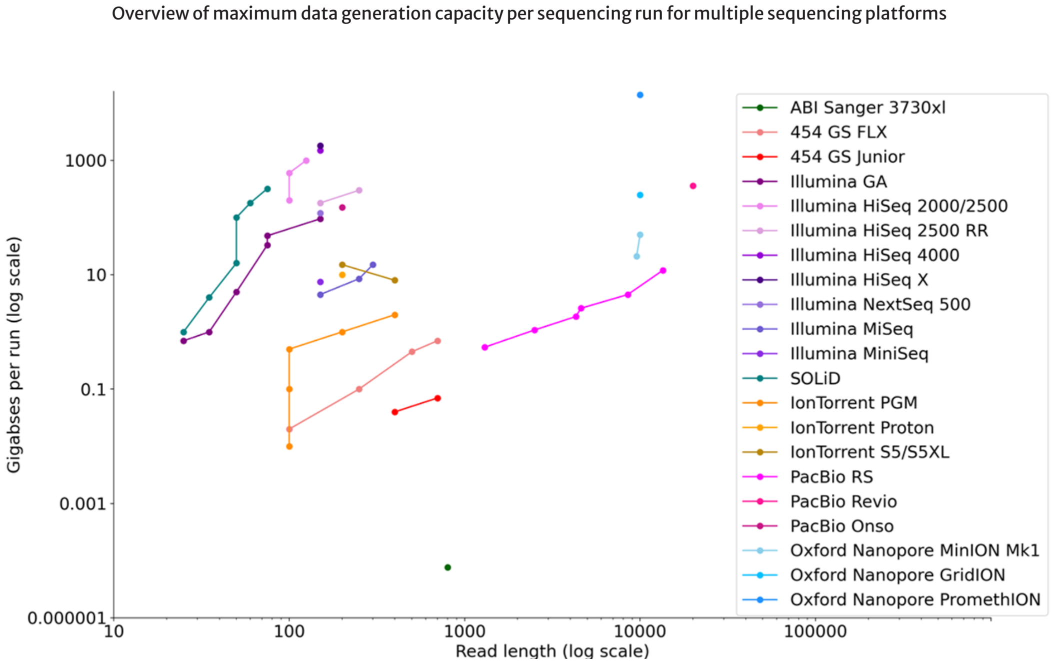

**Figure 7 Selection of sequencing devices and their corresponding maximum output per sequencing run.** Run-times are predefined by the sequencing platform and deviate from 24 to 72 hours. Platforms are visualized through color-coding; different versions of the same platform are depicted using different color shades. For truly large genomes with complex genomic structures currently only the PromethION device from Oxford Nanopore Technologies provides an adequate solution (single dot in upper-right corner). Figure adopted from[22].

## Downstream analysis – alignment and difficulties

The ability to generate dataset volumes that contain multiple full genome copies has been the focus of interest ever since development of sequencing platforms. However downstream analysis has become more difficult for datasets of increasing volume, both computationally as well as storage and distribution. Hence the general focus is moving away from simply generating larger sequencing datasets. Since generating a large data set requires adequate computational methods it has become increasingly evident that the development of algorithms need to allow analysis of large and complex genome datasets.

A simple workflow from sequencing data to the final *de novo* assembly involves:
1) read alignment, 2) assembly with read overlap graphs and 3) error correction (either performed at the start, middle or end of the assembly procedure). Starting off *de novo* assembly is alignments. Long-read alignment quickly becomes problematic since every read must be compared to every other read to find overlap for assembly[23]. The total number of comparisons is $n^2$, where n stands for the number of reads. This challenge becomes more prominent using larger datasets since the search-space increases quadratically. As an example, for *de novo* assembly we need multiple copies of the complete genome to allow reads to overlap. In a theoretical example we can use 50x coverage of a simple bacterial genome with an approximate genome size of 5 Mbp. Calculating the data volume needed for a 50x coverage of a 5 Mbp genome dataset yields a 50 x 5 = 250 Mbp dataset. When using long reads with approximately 10 Kbp read lengths this dataset comprises 25,000 reads. For *de novo* assembly every read is compared against every other read leading to $25{,}000^2$ = 625,000,000 comparisons for a dataset containing a simple bacterial genome. Then, if every read is 10 Kbp long and assuming optimal global alignments, pairwise comparison requires a 10,000 x 10,000 dynamic programming matrix to compare all bases and find the best alignment between the two reads. Hence $(6{,}2 \times 10^8) \times (1 \times 10^4) \times (1 \times 10^4) = 6{,}2 \times 10^{16}$ (62 quadrillion base comparisons) (Figure 8, bacteria). Although modern alignment algorithms are much more sophisticated compared to this exemplified brute force alignment strategy, it does clearly indicate the challenge modern alignment algorithms are facing.
For a diploid human genome (total genome size 6 Gbp), using similar data specifications, all reads vs. all reads alignment quickly skyrockets to a staggering $9 \times 10^{14}$ read comparisons, and $(9 \times 10^{14}) \times (1 \times 10^4) \times (1 \times 10^4) = 9 \times 10^{22}$ base comparisons using 30 million 10 Kbp reads (Figure 8, human).

Classical assembly strategies reconstructing the human genome have used thousands of CPU hours and over 100 Gb memory to finish the assembly of the human genome[10]. This implies that assemblies for truly large genomes, such as for plants that have high ploidy, require the entire lifespan of a specialized computer cluster to finish all comparisons (Figure 8, *Tulipa gesneriana*). Moreover, the lack of scalability puts a large pressure on both time and resources, making it unfeasible to reconstruct truly large genomes in a standardized fashion.

In chapter 2 and 4 we have proposed an alternative strategy for larger genomes, where we have decreased the search space by strategically selecting short sequences, reads restricted to a fixed size are called seeds. Those seeds can be selected from alternative sequencing technologies, such as Illumina reads, or randomly sampled from long read sequencing data. Applying this alternative method to the example above reveals the release of resource pressure during *de novo* assembly and the scalability towards truly large genomes (Figure 8, "reads vs seeds" in blue).

All reads vs all reads alignment search-space increases exponentially, hence computationally infeasible for truly large genomes. Restricting reference data towards coverage and read length reliefs resource pressure significantly

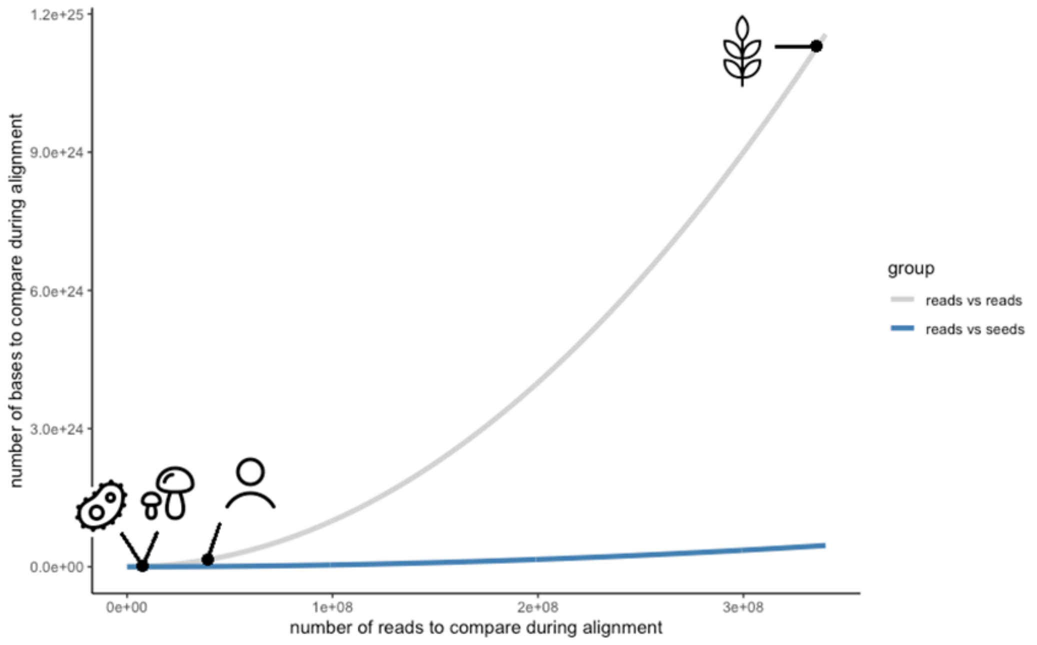

**Figure 8 Illustrating the exponential growth for all reads versus all reads using 50x coverage and 10 kbp reads.** Genomes from small to large; bacterial genome ~5 Mbp, fungal genome ~100 Mbp, human (diploid) genome ~6 Gbp, Tulipa gesneriana genome (diploid) ~68 Gbp. Restricting reference data transforms the exponential search-space-growth during alignment to a nearly linear fashion and highlights the scalability particularly for truly large genomes (in blue).

## Reference-based assembly, *de novo* and hybrid *de novo* assembly

Assembly has been a cornerstone in reconstructing amplicons and full genomes, the development of third generation sequencing data has been beneficial for, among others, identification of structural variation. The most straightforward assembly method is reference-based assembly, where the genomic architecture of the organism of interest has been reconstructed in previous studies. This reference assembly is then used as a guide to reconstruct the sequencing data set, allowing for a swift and relatively simple assembly task. However, most organisms that are of scientific interest do not have published assembly references and require a full reconstruction of the genetic code solely based on the sequencing data set.

Such assembly is referred to as *de novo* assembly. NGS *de novo* assemblies usually yield high quality assemblies due to the high per base quality of those platforms. However, those kinds of sequencing data lack the ability to span over large repetitive regions and hence usually result in highly fragmented assembly results. Therefore, combining both short and long-read sequencing data provides the best of both worlds, referred to as hybrid *de novo* assembly.

Hybrid *de novo* assembly on the one hand provides a high quality yet highly fragmented backbone assembly, and then uses the comprehensive structural information delivered by third generation sequencing data to bridge gaps caused by large repetitive regions. During the past decade numerous assembly tools have been developed, some focusing on structural correctness and contiguity, others on relieving computation pressure and down scaling required resources. A definite gold-standard workflow has not fully emerged. The current gold-standard in finding the correct assembly method is to perform multiple assemblies and compare assembly results in order to answer the research question at hand[13, 14, 24–26].

## Error correction

Resolving misclassified bases in the final genome assembly is another relatively resource intensive part of the assembly workflow for third generation sequencing data. A large variety of correction tools have been developed and essentially provide two flavors; first self-correction, where copies of long reads are aligned against each other and errors in a single copy are corrected using a majority vote. Those high quality long-reads are then used as input for *de novo* assembly and decrease the assembly complexity since the assembly graph does no longer contain split paths due to sequencing errors. PacBio yields HiFi reads based on this self-correction approach.

The second flavor uses data from multiple platforms, and similar to hybrid *de novo* assembly, combines the best of two worlds, long reads that provide structural information captured within the read itself can link contigs generated from NGS data. Those regions are then corrected based on the high quality of short-read sequencing data. Hybrid correction currently outperforms most self-correcting methods in terms of rescuing base misclassification. It is worth mentioning that hybrid correction usually requires lower memory resources despite higher CPU usage and requires the same organism to be sequenced through multiple platforms[11–12].

## The cost of genome sequencing

Starting off in the 90's the human genome project was one of the first large scale sequencing projects, aiming to fully reconstruct the human genetic code. It took approximately 13 years, thousands of researchers and roughly $3 billion to reconstruct the chromosomes. A decade ago, that price had dropped to $10K, and the last couple of years prices have decreased even further around $600 per human genome. The speeds at which sequencing cost is declining outpaces Moore's law, which states that the number of transistors on microchips double every two years, which translates to double computer power every two years. Technology improvements that follow the trend defined by Moore's Law are known to perform exceedingly well, Moore's Law is therefore a suitable comparison evaluating sequencing cost[33]. the National Human Genome Research Institute (NHGRI) has collected pricings for sequencing human genomes over time, see Figure 9. Here we clearly observe the drastic outpace of Moore's Law, particularly since 2008 where a prominent shift occurred moving away from Sanger sequencing and introducing NGS sequencing data technologies.

Comparison of well-defined Moore's law and sequencing cost, indicating the astonishing speed at which sequencing is evolving

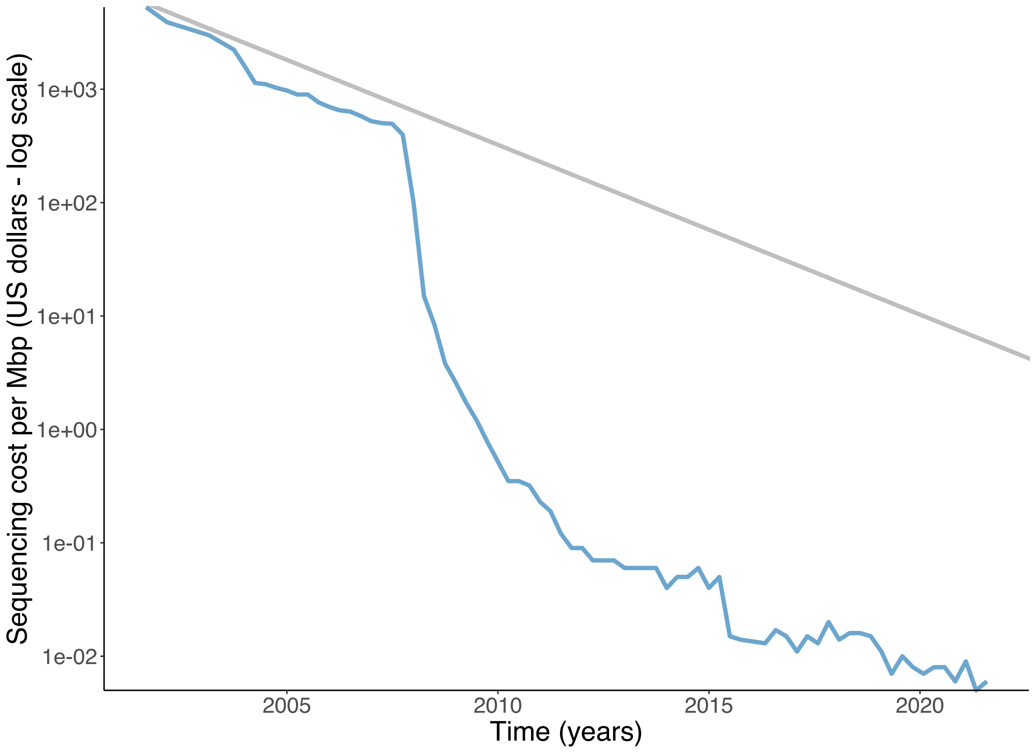

Figure 9 **Sequencing cost (in blue) compared to Moore's law (in gray), developing technologies that follow Moore's law are known to perform exceedingly well.** The outpacing of this law is a strong indication of the promising impact of sequencing technology and its applications. Figure adopted from[27].

To collect sequencing data researchers either send their isolated DNA to companies or obtain sequencing facilities inhouse. Using sequence service providers has been the gold-standard for many laboratories and institutes that do not have a core focus on genome analysis. The benefit of outsourcing sequencing projects is that per-base genome sequencing has become affordable and the per-base sequencing cost is now at the level of cents per base. Although it is true that sequencing cost has declined incredibly, there are considerable discrepancies in the total cost of genome sequencing, which stem from commercial purposes where media and even academic literature attempt to highlight technologies and platforms most opportunistically. A clear overview of the total sequencing expense therefore remains absent due to large variation between sequencing devices and release models, chemistries, bulk-ordering and sequencing service offers. However, to get an initial impression on the total sequencing cost the two most significant factors are equipment and chemistry (operation). The table below is a brief overview of those two costs. Price estimations are based on the most expensive variation reported, hence no bulk-ordering, special offers or discounts are taken into consideration.

Table 1 Two major factors for total sequencing cost, sequencing device (estimated regardless of machine type and release model) and chemistry (flow cells and library prep kits).

| | Illumina | PacBio | Oxford Nanopore Technologies |
|---|---|---|---|
| *Seq device* | ~ $100K – 1M | ~ $1M | ~ $2 – 50K |
| *consumables* | $6K | $4K | $1.5 -3K |

It also must be taken into consideration that to acquire sequencing devices, such as for Illumina and PacBio, additional environmental factors might have to be considered, such as temperature, humidity, air quality, cooling, and/or floor reinforcement. This could significantly contribute to the total cost for sequencing. ONT devices do not demand stringent environmental control requirements, only requiring temperatures between 18 and 24°C and thereby provide an advantage economically as well as facilitating user friendliness.

## References

1. Rboert W. Holley et al., (1965), Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid, The journal of biological chemistry, Vol. 240, No. 5

2. Marina Barba et al., (2014), Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology, 6, 106-136; doi:10.3390/v6010106

3. Fábio C. P. Navarro et al., (2019), Genomics and data science: an application within an umbrella, Genome Biology 20:109 https://doi.org/10.1186/s13059-019-1724-1

4. Zhong Wang et al., (2009), RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet.; 10(1): 57–63. doi:10.1038/nrg2484.

5. Udo Oppermann et al., (2013), Why is epigenetics important in understanding the pathogenesis of inflammatory musculoskeletal diseases?, Arthritis Research & Therapy volume 15, Article number: 209, https://doi.org/10.1186/ar4186

6. Shuikan, A. et al., (2019), High-Throughput Sequencing and Metagenomic Data Analysis. In (Ed.), Metagenomics - Basics, Methods and Applications. IntechOpen. https://doi.org/10.5772/intechopen.89944

7. Anthony Rhoads, Kin Fai Au, PacBio Sequencing and Its Applications, Genomics, Proteomics & Bioinformatics, Volume 13, Issue 5, 2015, Pages 278-289, ISSN 1672-0229, https://doi.org/10.1016/j.gpb.2015.08.002.

8. Jannes Spangenberg et al., Prediction of differential single nucleotide changes in the Oxford Nanopore Technologies sequencing signal of SARS-CoV-2 samples. doi: https://doi.org/10.1101/2023.03.17.533105

9. Louise Aigrain, Senior Staff Scientist in the DNA Pipelines Research and Development team - at the Wellcome Sanger Institute. Image credit: Genome Research Limited. https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

10. Jain, M. et al., (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature biotechnology, 36(4), 338–345. https://doi.org/10.1038/nbt.4060

11. Zhang, H. et al., (2020), A comprehensive evaluation of long read error correction methods. BMC Genomics 21 (Suppl 6), 889. https://doi.org/10.1186/s12864-020-07227-0

12. Pierre Morisse et al., Long-read error correction: a survey and qualitative comparison, biorivx, doi: https://doi.org/10.1101/2020.03.06.977975

13. Boostrom et al., (2022). Comparing Long-Read Assemblers to Explore the Potential of a Sustainable Low-Cost, Low-Infrastructure Approach to Sequence Antimicrobial Resistant Bacteria With Oxford Nanopore Sequencing. Frontiers in microbiology, 13, 796465. https://doi.org/10.3389/fmicb.2022.796465

14. Khrenova et al., Nanopore Sequencing for *De Novo* Bacterial Genome Assembly and Search for Single- Nucleotide Polymorphism. Int. J. Mol.Sci.2022,23,8569. https:// doi.org/10.3390/ijms23158569

15. Zhenyu Li et al. (2012), Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Briefings in Functional Genomics, Volume 11, Issue 1, Pages 25–37, https://doi.org/10.1093/bfgp/elr035

16. Pevzner et al., (2001). An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America, 98(17), 9748–9753. https://doi.org/10.1073/pnas.171285098

17. Robert M Waterhouse et al., (2018), BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics, Molecular Biology and Evolution, Volume 35, Issue 3, Pages 543–548, https://doi.org/10.1093/molbev/msx319

18. Wang, T., Antonacci-Fulton, L., Howe, K. et al., (2022), The Human Pangenome Project: a global resource to map genomic diversity. Nature 604, 437–446. https://doi.org/10.1038/s41586-022-04601-8

19. Attwaters, M., (2022), The final pieces of the human genome. Nat Rev Genet 23, 321.
https://doi.org/10.1038/s41576-022-00494-5

20. Rautiainen, M., Nurk, S., Walenz, B.P. et al., (2023), Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. https://doi.org/10.1038/s41587-023-01662-6

21. Browne, Patrick Denis et al., GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms, GigaScience, DO 10.1093/gigascience/giaa008, https://doi.org/10.1093/gigascience/giaa008

22. https://dx.doi.org/10.6084/m9.figshare.100940

23. Myers E. W. (2005). The fragment assembly string graph. Bioinformatics (Oxford, England), 21 Suppl 2, ii79–ii85.
https://doi.org/10.1093/bioinformatics/bti1114

24. Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome research, 27(5), 737–746. https://doi.org/10.1101/gr.214270.116

25. Jain, M., (2018), Nanopore sequencing and assembly of a human genome with ultra-long reads.
Nature biotechnology, 36(4), 338–345. https://doi.org/10.1038/nbt.4060

26. Wick RR and Holt KE., (2021), Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 4; peer review: 4 approved]. F1000Research, 8:2138 (https://doi.org/10.12688/f1000research.21782.4)

27. Kris A. Wetterstrand, . https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

28. Payne A et al., (2018), Whale watching with BulkVis: a graphical viewer for Oxford Nanopore bulk fast5 files.
BioRxiv https://doi.org/10.1101/312256

29. Gregory W Vurture et al., (2017), GenomeScope: fast reference-free genome profiling from short reads, Bioinformatics, Volume 33, Issue 14, July 2017, Pages 2202–2204, https://doi.org/10.1093/bioinformatics/btx153

30. International Human Genome Sequencing Consortium. (2001), Initial sequencing and analysis of the human genome.
Nature 409, 860–921. https://doi.org/10.1038/35057062

31. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849–64.

32. Yang, X., Lee, WP., Ye, K. et al., (2019), One reference genome is not enough. Genome Biol 20, 104.
https://doi.org/10.1186/s13059-019-1717-0

33. Moore, G., (1965), Moore’s law. Electronics Magazine, 38(8), 114.

## — Thesis outline

In this thesis I evaluate what the impact of Oxford Nanopore Technologies on different genome sequencing applications is. I aimed to investigate whether the distinctive properties of long-read sequencing can be useful for several sequencing applications, with an emphasis on downstream bioinformatics analyses.

### Chapter 2

We ask whether there is a large impact of the development of the ONT platform on genome assemblies using a wild yeast isolate. This yeast isolate contains GC-biased regions which make it difficult for sequencing technologies to generate homogeneous coverage. Additionally, this genome contains abundantly present duplications and repetitive content that yields fragmented assemblies using NGS data alone. We therefore additionally utilize long read sequencing data to close gaps and to resolve repetitive content.

### Chapter 3

Here we investigate the genomic structure of an *Allorhizobium* strain (LBA9072) that was formerly considered *Agrobacterium*. *Agrobacterium* genomic architecture is usually of moderate complexity, including a couple of large chromosomes, a tumor-inducing plasmid, and additional plasmids of unknown function. *Agrobacterium* is extensively used for genetic engineering since this plant-pathogen can transfer and integrate parts of its own genome into the genome of the host plant. The genomic architecture of this strain contains many repetitive sequences; hence it deviates from most well-studied *Agrobacterium* strains. Those repetitive regions are larger than the longest NGS reads, additionally the architecture of microbial genomes is very dynamic between related strains and species. Both arguments suggest that long-read sequencing data might be needed to fully reconstruct the genomic content.

## Chapter 4

In this chapter we ask whether the MinION can be used to sequence large eukaryotic genomes. The throughput of the MinION has increased rapidly since its introduction, making this a possibility in theory. However, the data characteristics and quality require a rethink of the *de novo* assembly process. We have sequenced the genome of the endangered European eel (medium-large genome size) and assembled those data with a novel assembly algorithm that is tailored for large eukaryotic genomes. We have evaluated the assembly algorithm performance and compared total assembly length and structural correctness to the draft assembly genome generated using short reads alone. Additionally, we evaluated genome completeness, contiguity, and structural correctness, as well as the computational resources reuired.

## Chapter 5

In this study we evaluate the utility of ONT sequencing to characterize microbial diversity in seawater from multiple locations. We test whether it is possible to characterize organisms at the species level, and if it supports the reconstruction of large contigs or entire genomes. We aimed to establish an initial workflow for environmental samples to assess the portable characteristics of the MinION device, using minimal resources and computing capacity.

## Chapter 6

Finally, I summarize and discuss the preceding chapters, and highlight the overall conclusions regarding the applicability of ONT sequencing data towards large and complex genome reconstruction. Furthermore, I glance into future applications of long-read sequencing.

— Chapter 2

# *De novo* whole-genome assembly of a wild type yeast isolate using nanopore sequencing

F1000Research 2018, 6:618

Michael Liem[1], Hans J. Jansen[2], Ron P. Dirks[2], Christiaan V. Henkel[1], G. Paul H. van Heusden[1], Richard J.L.F. Lemmers[3], Trifa Omer[4], Shuai Shao[1], Peter J. Punt[1,4], Herman P. Spaink[1]

[1] Institute of Biology, Leiden University, Leiden, 2300 RA, Netherlands

[2] Future Genomics Technologies B.V., Leiden, 2333 BE, Netherlands

[3] Department of Human Genetics, Leiden University Medical Center, Leiden, 2333 ZA, Netherlands

[4] Dutch DNA Biotech B.V., Utrecht, 3584 CH, Netherlands

## — Abstract

### Background

The introduction of the MinION sequencing device by Oxford Nanopore Technologies may greatly accelerate whole genome sequencing. Nanopore sequence data offers great potential for *de novo* assembly of complex genomes without using other technologies. Furthermore, Nanopore data combined with other sequencing technologies is highly useful for accurate annotation of all genes in the genome. In this manuscript we used nanopore sequencing as a tool to classify yeast strains.

### Methods

We compared various technical and software developments for the nanopore sequencing protocol, showing that the R9 chemistry is, as predicted, higher in quality than R7.3 chemistry. The R9 chemistry is an essential improvement for assembly of the extremely AT-rich mitochondrial genome. We double corrected assemblies from four different assemblers with PILON and assessed sequence correctness before and after PILON correction with a set of 290 Fungi genes using BUSCO.

### Results

In this study, we used this new technology to sequence and *de novo* assemble the genome of a recently isolated ethanologenic yeast strain, and compared the results with those obtained by classical Illumina short read sequencing. This strain was originally named *Candida vartiovaarae* (*Torulopsis vartiovaarae*) based on ribosomal RNA sequencing. We show that the assembly using nanopore data is much more contiguous than the assembly using short read data. We also compared various technical and software developments for the nanopore sequencing protocol, showing that nanopore-derived assemblies provide the highest contiguity.

### Conclusions

The mitochondrial and chromosomal genome sequences showed that our strain is clearly distinct from other yeast taxons and most closely related to published *Cyberlindnera* species. In conclusion, MinION-mediated long read sequencing can be used for high quality de novo assembly of new eukaryotic microbial genomes.

Keywords Nanopore sequencing, *de novo* genome assembly, wild type yeasts, ethanologenic, *Candida*, *Cyberlindera*

## — Introduction

With the development of robust second generation bioethanolprocesses, next to the use of highly engineered *Saccharomyces cerevisiae* strains[1,2], non-classical ethanologenic yeasts are also being considered as production organisms[3,4]. In particular, aspects concerning the ability to use both C6 and C5 C-sources and feedstock derived inhibitor resistance have been identified as important for the industrial applicability of different production hosts[3]. In our previous studies we have identified a novel ethanologenic yeast, *Wickerhamomyces anomala*, as a potential candidate[3]. Based on this research, a further screen for alternative yeast species was initiated (Punt and Omer, unpublished study). Here we describe the isolation and genomic characterization of one of these new isolates, which was typed as *Candida vartiovaarae* based on ribosomal RNA analysis.

With the arrival of next generation sequencing and the assemblers that can use this type of sequencing data, whole genome shotgun sequencing of completely novel organisms has become affordable and accessible. As a result, a wealth of genomic information has become available to the scientific community leading to many important discoveries. While generating whole draft genomes has become accessible, these genomes are often fragmented due to the nature of these short read technologies[5]. Assembling short read data into large contigs proved to be difficult because the short reads do not contain the information to span repeated structures in the genome. Approaches to sequence the ends of larger fragments partially mitigated this problem[6].

The new long read platforms from Pacific Biosciences and Oxford Nanopore Technologies made it possible to obtain reads that span many kilobases[7]. Assemblies using this type of data are often more contiguous than assemblies based on short read data[8,9].

We have employed the Oxford Nanopore Technologies MinION device to sequence genomic DNA from the isolated *Candida vartiovaarae* strain. The same DNA was also used to prepare a paired end library for sequencing on the Illumina HiSeq2500. The sequence data were used in various assemblers to obtain the best assemblies.

## — Materials and methods

### Strain selection and cultivation conditions

In our previous research[3], a screening approach was developed to select for potential ethanologens using selective growth on industrial feedstock hydrolysates. Based on this approach, a previously identified microflora from grass silage was screened for growth on different hydrolysates from both woody and cereal residues. From this microflora, a strain was isolated (DDNA#1) after selection on a growth medium consisting of 10% acidpretreated corn stover hydrolysate, which was shown to be most restrictive in growth due to the presence of relatively high amounts of furanic inhibitors.

### DNA purification

Cells were grown at 30°C on plates with YNB (without amino acids) medium supplemented with 0.5% glucose. Cells were scraped from plates and resuspended in 5 ml TE. High MW chromosomal DNA was isolated from yeast isolate DDNA#1 and *Saccharomyces cerevisiae* S288C using a Qiagen Genomic-tip 100/G column, according to the manufacturer's instructions.

### Pulsed field gel electrophoresis

In order to determine the size of intact chromosomes of DDNA#1, a BioRad CHEF Genomic DNA Plug Kit was used. Briefly, yeast cells were treated with lyticase and the resulting spheroplasts were embedded in low melting point agarose. After incubation with RNase A and Proteinase K, the agarose plugs were thoroughly washed in TE. The DNA in the agarose plugs was separated on a 0.88% agarose gel in 1xTAE buffer on a Bio-Rad CHEF DRII system. The DNA was separated in four subsequent 12 hour runs at 3V/cm; run one and two used a constant switching time of 500 seconds, and in run three and four the switching time increased from 60 seconds to 120 seconds. The gel was afterwards stained with ethidium bromide and imaged.

### Genome size estimation and heterozygosity

A k-mer count analysis was done using Jellyfish[10] v2.2.6 on the Illumina data. From the paired end reads, only the first read was truncated to 100 bp to avoid the lower quality part of the read. The second read was omitted from this analysis to avoid counting overlapping k-mers. Different k-mer sizes were used ranging from k=17 to 23. After converting the k-mer counts into a histogram format, this file was analyzed using the Genomescope[11] tool, available at http://qb.cshl.edu/genomescope/ and https://github.com/schatzlab/genomescope.

### Illumina library preparation, sequencing and quality control

High molecular weight DNA from both DDNA#1 and *Saccharomyces cerevisiae* S288C was sheared using a nebulizer (Life Technologies). The sheared DNA was used to make genomic DNA libraries using the Truseq DNA sample preparation kit, according to the manufacturer's instructions (Illumina Inc.). In the size selection step, a band of 330–350 bp was cut out of the gel to obtain an insert length of ~270 bp. From the resulting libraries, 4.5 million fragments were sequenced in paired end reads with a read length of 150 nt on an Illumina HiSeq2500, according to the manufacturer's instructions.

The HiSeq control software (HCS) and real time analysis (RTA) software, versions were 2.2.38 and 1.18.61, respectively, were used. To ensure data integrity we have visualized read quality distributions with FastQC[12] v0.11.7 and merged overlapping paired end reads, including trimming of low quality regions, using flash[13] v1.2.11. Only trimmed and merged reads are used as input data for both Spades[14] assemblies and assembly polishing.

## MinION library preparation, sequencing and quality control

The genomic DNA was sequenced using nanopore sequencing technology. First the DNA was sequenced on R7.3 flow cells. Subsequently, multiple R9 and R9.4 flow cells were used to sequence the DNA. For R7.3 sequencing runs, we prepared the library using the SQK-MAP006 kit from Oxford NanoporeTechnologies. In short, high molecular weight DNA was sheared with a g-TUBE (Covaris) to an average fragment length of 20 kbp. The sheared DNA was repaired using the FFPE Repair Mix, according to the manufacturer's instructions (New England Biolabs). After cleaning the DNA with bead extraction, using a ratio of 0.4:1 Ampure XP beads (Beckman Coulter) to DNA, the DNA ends were polished and an A overhang was added with the NEBNext End Prep Module (New England Biolabs).
Then, prior to ligation, the DNA was again cleaned by extraction using a ratio of 1:1 Ampure XP beads to DNA. The adaptor and hairpin adapter were ligated using Blunt/TA Ligase Master Mix (New England Biolabs). The final library was prepared by cleaning the ligation mix using MyOne C1 beads (Invitrogen).

To prepare 2D libraries for R9 sequencing runs, we used the SQK-NSK007 kit from Oxford Nanopore Technologies. The procedure to prepare a library with this kit is largely the same as with the SQK-MAP006 kit. 1D library preparation was done with the SQK-RAD001 kit from Oxford Nanopore Technologies, which tags high molecular weight DNA using a transposase. The final library was prepared by ligation of the sequencing adapters to the tagmented fragments using the Blunt/TA Ligase Master Mix (New England Biolabs).

The prepared libraries were loaded on the MinION flow cell, which was docked on the MinION device. The MinKNOW software (v0.50.2.15 for SQK-MAP006 libraries and v1.0.5 for SQK-NSK007 and SQK-RAD001 libraries) was used to control the sequencing process and the read files were uploaded to the cloud based Metrichor EPI2ME platform for base calling. Base called reads were downloaded in fastq format. We filtered the data to a per read average maximum error-rate distribution of 10% and a minimum of 10 kbp for quality and length, respectively. Only reads that meet these filtering thresholds were used for assemblies and post-assembly error correction.

## Genome assembly and assembly correction

The sequence data from the Illumina platform was assembled using Spades v3.6.0, we performed a two-branch assembly strategy using either exclusively Illumina data or a hybrid approach combining both Illumina and nanopore data sets.

A set of four different assemblers is used to generate contigs exclusively based on nanopore data, Canu[15] v1.3, Miniasm[16] v0.2, TULIP[17] v0.4 and Smartdenovo18 v1.07. These assemblers perform all vs. all alignments on filtered nanopore data to generate the final contigs, with the exception of TULIP, which aligns reads to a set of random 1,000 bp seed sequences comprising 0.5 times the estimated ~12 Mbp genome size. Contigs of all assemblers were post-assembly corrected using Racon[19], excluding Canu generated contigs, since Canu contains an integrated self-correction procedure prior to assembly. To obtain optimum sequence correctness the resulting contigs of these four assemblers were polished with Illumina data using PILON[20] v1.18 in a double iterative fashion. The sequencing data, including the final assembly, has been submitted to the European Nucleotide Archive and can be accessed at http://www.ebi.ac.uk/ena/data/view/PRJEB19912.

## Genome assembly assessment based on gene prediction

As successful sequence polishing plausibly improves the accuracy of gene prediction, we assessed both assembly quality and PILON correction effects using BUSCO[21] v3.0.2. We assessed our nanopore exclusive assemblies both before and after PILON correction using lineage database Fungi 0db9 containing 290 genes. BUSCO genome assembly assessments on Spades contigs correspond to assessments after PILON correction for nanopore derived contigs, since Spades contigs are based on Illumina data and do not require a post-assembly PILON correction. BUSCO identifies genes in genomic assemblies either as partial, single or double copy, or completely absent.

## Full genome comparison

From 26S ribosomal RNA sequences available in the nucleotide database, Chen *et al.*[22] have constructed a phylogenetic tree. From that phylogenetic tree we have observed that the closest relative for which whole genome sequences are available is *Cyberlindnera jadinii*. To compare our draft genome assembly to this yeast species, we retrieved assemblies of two *Cyberlindnera jadinii* strains, namely NBRC 0988 (GenBank accession number, DG000077.1) and CBS1600 (GenBank accession number, CDQK00000000.1). We also used *Saccharomyce cerevisiae* S288C (GenBank accession number, GCA_ 000146045.2) in this comparison. We aligned those assemblies to the corrected draft assembly of our strain using MUMmer's alignment generator NUCmer[23] v3.1). NUCmer's output was filtered and the filtered results parsed to MUMmerplot, generating full-genome visualization between the pairs of different yeast species. Since Spades assembly-lengths are roughly twice the estimated genome size we additionally evaluated alignments between Spades hybrid and TULIP contigs. Alignments were performed using BWA-mem[24] v0.7.15 with -x ontd2 settings and visualized using genome viewer Tablet[25] v1.17.08.17.

## Read mapping to mitochondrial genome

Reads generated on the Illumina platform were aligned to the published *Candida vartiovaarae* mitochondrial genome (Genbank accession number, KC993190.1) using Bowtie2[26] v2.2.5. Reads generated on the MinION platform were aligned using Minimap2[27] v2.3-r546-dirty. Resulting bam files were sorted and viewed in IGV viewer v2.3.

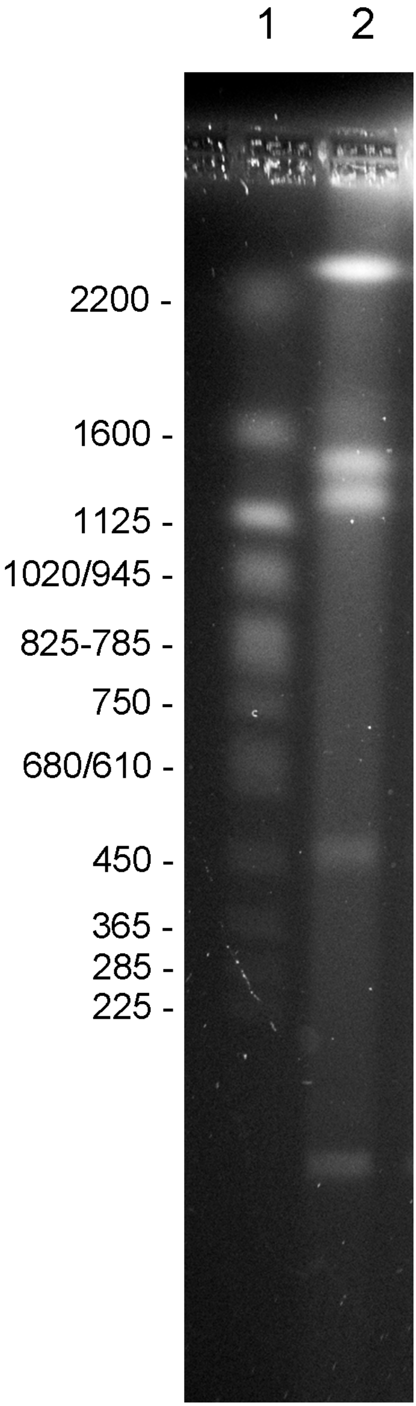

Figure 1 **Pulsed field gel electrophoresis of *Candida vartiovaarae* DDNA#1 chromosomes.**
In lane 1, the chromosomes of Saccharomyces cerevisiae were loaded as a marker. Sizes of the chromosomes in the marker lane are indicated. In lane 2, thechromosomes of *Candida vartiovaarae* DDNA#1 were loaded.

## — Results and discussion

### Pure cultures of candidate ethanologenic yeasts

From a screen on 10% acid-pretreated corn stover hydrolysate, about 70 individual clones were obtained, only five of which were able to grow well on purely synthetic YNB-based medium. To determine the taxonomic status of these clones, chromosomal DNA was isolated and used for PCR amplification of the ribosomal ITS sequence using ITS-specific primers[28] (ITS1 andITS4). BLAST analysis of these ITS sequences of all 5 isolates revealed a 100% identity to *Candida vartiovaarae* (*Torulopsis vartiovaarae*: NCBI accession number KY102493)

All five isolates were grown on different C-sources and showed growth on glucose, mannose, cellobiose, xylose and glycerol, while growth on L-arabinose was variable. No significant growth was found on galactose and rhamnose. Good growth (on glucose) occurred between 20–30°C, at pH3-7 (optimum 25°C, pH4-5). Based on the results, we concluded that all five isolates originated from a single source in the grass silage sample. Subsequent experiments were therefore carried out with a single isolate now named DDNA#1.

### Pulsed field gel electrophoresis

As a further means to validate our assembled contigs and determine if they match the actual chromosome length, we have separated the chromosomes on an agarose gel using pulsed field gel electrophoresis. The gel image in Figure 1 shows five bands that represent the chromosomes of this yeast strain. The smallest band has a length that corresponds to the length of the mitochondrial genome (33 kbp). Additional fragments of 450, 1200, and 1500 kbp are also found. The intensity of the band that runs above the 2200 kbp marker band suggests that it actually contains more than one distinct fragment. To make the genome size fit to the estimate derived from the assembly and k-mer analysis (~12.5 Mbp), three ~3 Mbp chromosomes should be postulated. The uncertainty in chromosome size estimate based on pulsed field electrophoresis gels is high because of the large chromosome size and the fact that it is difficult to determine if more than one fragment is present in the gel at a given position. Our conclusion that the top band represents three or more chromosomes is in agreement with the genome sequences of two related *C. jadinii* strains, namely CBS1600 and NBRC 0988.

## Genome size estimation and heterozygosity

The Illumina sequence data of our DDNA#1 isolate were submitted to the Genomescope software package to analyze the k-mer count distribution, using k-mer size = 19 at an average coverage of 28.0x (Figure 2). The 'haploid' genome is predicted to contribute to the most abundant fraction, which corresponds with the second peak (dotted line) in the plot (Figure 2). The first peak corresponds to sequence occurring exactly half as frequently as the main peak, so these are plausibly haplotypes. Due to the nature of k-mer counting, this peak often appears higher than the main peak, because a single SNP will affect all k-mers overlapping that position. The first two peaks contain about 10 Mbp of sequence. Additional peaks at higher coverage indicate duplications and repetitive DNA that are quite abundant, but correspond with less sequence than the second peak. Genomescope estimated a haploid genome size of between 12.00 and 12.01 Mbp. Additionally, Genomescope revealed 3.6% variety across the entire genome indicating that the genome of *C. vartiovaarae* has strong heterozygous properties (Table 1). A likely possibility is that areas in the genome are replicated and slightly diverged in sequence. This could also explain why we see a large tail of repeated k-mers (Figure 2). It could also explain why our assembly still remained fragmented despite the relatively large amount of nanopore data that was used in the assembly.

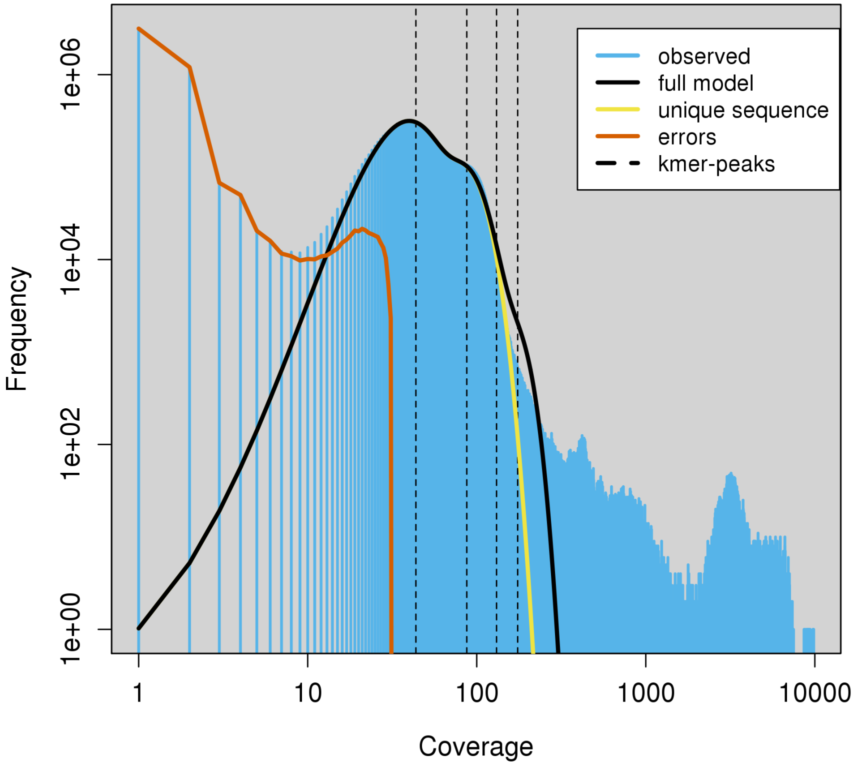

Figure 2 **Genome size estimation generated by Genomescope, providing a k-mer analysis (k = 19, from Jellyfish) to estimate haploid genome size, fraction of heterozygosity and coverage.**
Genomescope attempts to find k-mer count peaks, low and high coverage peaks indicating hetero- and homozygosity. (A) We find ~13× and ~28× coverage for hetero- and homozygous fractions in our dataset. Exact peakpositions are determined with a log transformation. Evaluating the slope between coverage points reveals the peak positions indicating hetero- and homozygosity, for lower and higher coverage, respectively.

Table 1 **Most important metrics from Genomescope.**

| k = 19 | k-mer coverage | 28.0 |
|---|---|---|
| property | min | max |
| Heterozygosity (%) | 3.64 | 3.65 |
| Genome Haploid Length (bp) | 11,995,570 | 12,010,675 |
| Genome Repeat Length (bp) | 2,179,917 | 2,182,662 |
| Genome Unique Length (bp) | 9,815,653 | 9,828,014 |
| Model Fit (%) | 98.26 | 98.89 |
| Read Error Rate (%) | 0.13 | 0.13 |

## Illumina and MinION *de novo* genome assembly

We took six approaches to assemble the genome of DDNA#1, five assemblies based on sequencing data from a single platform (either Illumina or nanopore) and one hybrid assembly. The first approach used reads exclusively produced by the Illumina platform. After merging paired end reads we obtained ~1.7 Gbp of ~240 bp reads. Contigs generated by Spades remained short and the overall assembly was heavily fragmented. The N50 of this assembly was only ~4.3 kbp, its longest contig ~35 kbp. Spades generated 10,121 contigs and the entire assembly length was nearly twice the estimated ~12 Mbp haploid genome size. We also assembled *Saccharomyces cerevisiae* S288 C using a similar short read dataset that was made and sequenced in parallel. Here we obtained an assembly that consisted of 768 contigs with a longer N50 of 124 kbp.

Assembly comparison of *Saccharomyces cerevisiae* and DDNA#1 exclusively based on Illumina data highlights that Spades clearly struggles to reconstruct the genome of our isolate, possibly due to complex SNP arrangements. From these results we take that, even under high coverage conditions, ~240 bp reads do not provide sufficient power to resolve complex SNP distributions for highly heterozygous genomes. This illustrates the necessity of increased read length to fully reconstruct complex genomic structures such as those found in DDNA#1.

Secondly, we used Spades to generate a hybrid assembly that takes both Illumina and nanopore data as input. We used ~1.7 Gbp and ~208 Mbp Illumina and nanopore data sets, respectively. This hybrid approach performed by Spades resulted in an N50 of ~379 kbp, with the longest contig ~1.1 Mbp, and a total of 653 contigs and, although still relatively fragmented, it is interesting that it yielded a similar assembly length compared to the assembly exclusively based on Illumina data. The improvement of assembly statistics strongly indicates the positive effect of longer reads in resolving complicated genomes.

Hereafter, the four remaining approaches are all based on data solely generated by the Oxford Nanopore Technologies platform. Assembly lengths in particular are fairly similar between all four assemblies and all approximate the estimated ~12 Mbp haploid genome size. However, Miniasm, TULIP and Smartdenovo outperform Canu on N50, number of contigs and longest contig (Table 2). Lengths of the longest contig from both Smartdenovo and TULIP (~2,8 Mbp) corresponds to the suggestion of ~3 Mbp chromosomes shown using pulse field gel electrophoresis on intact chromosomal DNA (Figure 1). This suggests that both Smartdenovo and TULIP were able to fully reconstruct one of the three largest chromosomes of our isolate. Although Smartdenovo results the lowest number of contigs, which is mainly due to a filtering step that filters out very short contigs (shortest contig lengths 1,716 bp and 73,332 bp for TULIP and Smartdenovo, respectively), TULIP generates the highest contiguity with N25 and N50 both around 1.6 Mbp compared to Smartdenovo that results in 1.4 Mbp and 900 kbp, respectively. Hence based on contiguity we prefer to take the TULIP result as the final assembly.

It is clear from these results that assemblies based on exclusively nanopore data achieve the most contiguous assemblies, as has been shown previously[8,9].

Table 2 Data characteristics and assembly statistics.

| Assemblers | Canu | Miniasm | TULIP | Smartdenovo | Spades hybrid | Spades |
|---|---|---|---|---|---|---|
| Data type | ONT | ONT | ONT | ONT | ONT and Illumina | Illumina |
| Reads (#) | 11,344 | 11,344 | 11,344 | 11,344 | 11,344 | 8,628,787 |
| Coverage (x) | 17 | 17 | 17 | 17 | 17 | 135 |
| GC-cont (%) | 46 | 46 | 46 | 46 | 46 | 47 |
| Bases (#) | 208,357,153 | 208,357,153 | 208,357,153 | 208,357,153 | 208,357,153 | 1,688,824,952 |
| Contigs | 34 | 25 | 28 | 20 | 653 | 10.121 |
| Assembly length (bp) | 11,968,989 | 12,072,133 | 11,325,084 | 11,732,656 | 22,772,746 | 22,356,011 |
| Genome size (Mbp) | 12.5 | 12.5 | 12.5 | 12.5 | 12.5 | 12.5 |
| N25 (bp) | 959,647 | 1,361,451 | 1,591,600 | 1,429,838 | 824,043 | 7,876 |
| N50 (bp) | 805,206 | 1,020,131 | 1,586,208 | 902,730 | 379,588 | 4,318 |
| N75 (bp) | 456,000 | 506,710 | 619,623 | 456,270 | 200,675 | 2,041 |
| Max length (bp) | 1,430,409 | 1,569,347 | 2,792,203 | 2,800,024 | 1,101,756 | 34,707 |
| Mean length (bp) | 352,029 | 482,885 | 404,467 | 586,632 | 34,874 | 2,208 |
| Min length (bp) | 4,727 | 8,316 | 1,716 | 73,332 | 128 | 128 |

We also used the nanopore datasets made with the R7.3 and R9 chemistry separately in the Canu assembler. The most notable difference between these assemblies is found in the mitochondrial genome. Only 16 kbp of this 33 kbp genome could be assembled with the R7.3 data, whereas the R9 assembly contained a complete mitochondrial genome (Genbank accession number, KC993190.1). The mitochondrial genome has a very low GC content (21%) and in the extragenic regions more A and T homopolymers are found. Very few R7.3 reads mapped to this region, but in the R9 dataset there are many more reads that represent this region (Figure 3). It has been shown that the R7.3 data especially has a bias against A and T homopolymers. Although this bias is still not fully absent[29,30], it is reduced for R9 chemestry, indicating technical enhancement and suggesting improved genomic reconstruction even for low complexity regions. Both after long read self-correction using Canu as well as for post-asssembly correction using Racon the contig sequences still contain errors[15]. We have used PILON and the complementary Illumina data from this strain to correct the assembled contigs twice. Homopolymer streches are paricularly difficult to base call accurately due to low complexity and lengths are usually underestimated. PILON correction leads to a minor assembly length increase since corrected homopolymer lengths adds to the final assembly size.

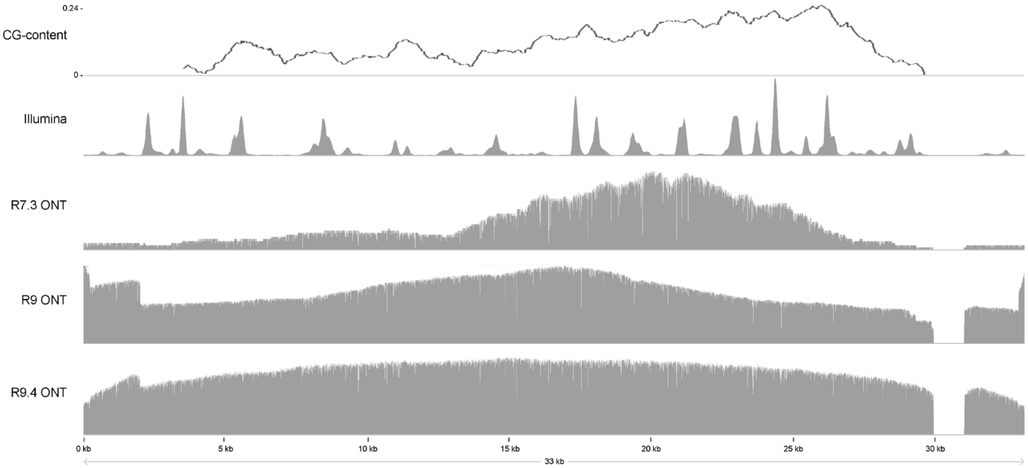

**Figure 3 Coverage plot of the *Candida vartiovaarae* DDNA#1 mitochondrial genome.**
Reads from both the Illumina, and the nanopore platform were aligned to the *Candida vartiovaarae* mitochondrial genome (Genbank accession number, KC993190.1) to show the difference in coverage between the different platforms and chemistry versions.

## Genome assembly assessment based on gene prediction

BUSCO identifies the majority of genes from database Fungi 0db9 on nanopore derived assemblies. The number of single copy genes identified ranges from 145 to 188, between 45 and 57 genes are partially recognized, and 53 to 92 genes are classified absent before PILON correction (Figure 4). After PILON correction nearly all genes are identified as single copies in the results from all four assemblers, giving support for the suggestion (based on genome size) that these assemblers yielded haploid genomes. Interestingly, gene identification on Spades contigs, particularly for our hybrid assembly, identified 269 genes as double copy genes. Together with assembly lengths of twice the estimated genome size these results strongly suggest that Spades was able to separately assemble both haplotypes forming a diploid genome under hybrid conditions. Only 100 and 67 genes are identified as double and single copy genes, respectively, for the Illumina exclusive assembly, again indicating the necessity of long read data to maximally reconstruct highly heterozygous genomes.

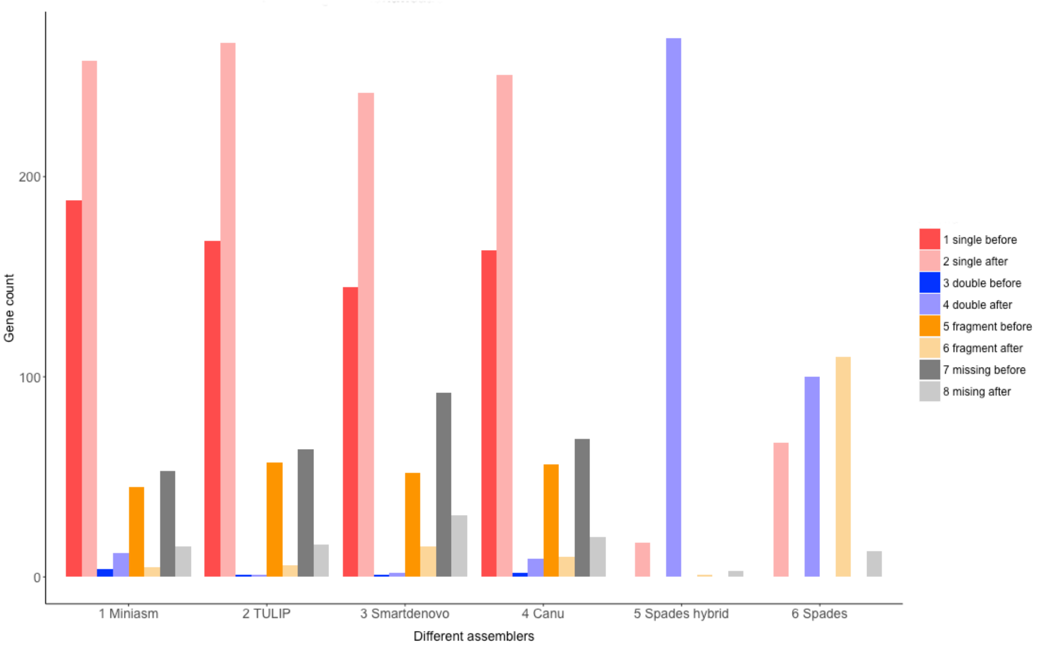

Figure 4 **BUSCO genomic assembly assessment using Fungi 0db9 database.**
Shown on the X-axis are 5 different assembler used in this study, including a hybrid assembly approach performed by Spades. Shown on the Y-axis are the Fungi 0db9 gene counts identified by BUSCO. Dark and light coloring shades indicate before and after PILON correction per classification type, respectively.

## Genome comparison

We have compared the assembled contigs of our *C. vartiovaarae* isolate DDNA#1 strain to yeast genome sequences that are already deposited in the nucleotide database. Comparison of our yeast strain with the well characterized *S. cerevisiae* assembly showed negligible genomic similarity. From 26S ribosomal RNA sequences available in the nucleotide database, Chen et al.[22] have constructed a phylogenetic tree. The closest relatives for which whole genome sequences are available are *C. jadinii* strains CBS1600 and NBRC 0988.
An initial comparison between CBS1600 and NBRC 0988 revealed that these two strains show high homology (Figure 5A). The genomic similarity between our strain and *C. jadinii* strains CBS1600 and NBRC 0988 is much lower (Figure 5B and Figure 5C, respectively). Assemblies exclusively based on nanopore data compared to Spades hybrid assembly strongly suggests the diploid properties of our strain, at least to a partial extend. At nearly every position on >90% of the TULIP assembly length a Spades hybrid contig is aligned. Figure 6 shows the longest TULIP contig and the third longest TULIP contig, ~2.9 and ~1.6 Mbp, respectively, and alignment of all possible Spades hybrid contigs. For TULIP contigs sorted on length we observe this double coverage behavior for contigs down to ~84 kbp. Shorter TULIP contigs tend to be less consistently double covered or even lack coverage of a Spades hybrid contig all together.
In conclusion, these data show that wild type yeast strains are very heterogeneous, despite a high similarity based on ribosomal RNA ITS sequences. Therefore, the data suggest that nanopore sequencing is an essential new tool to classify yeast strains.

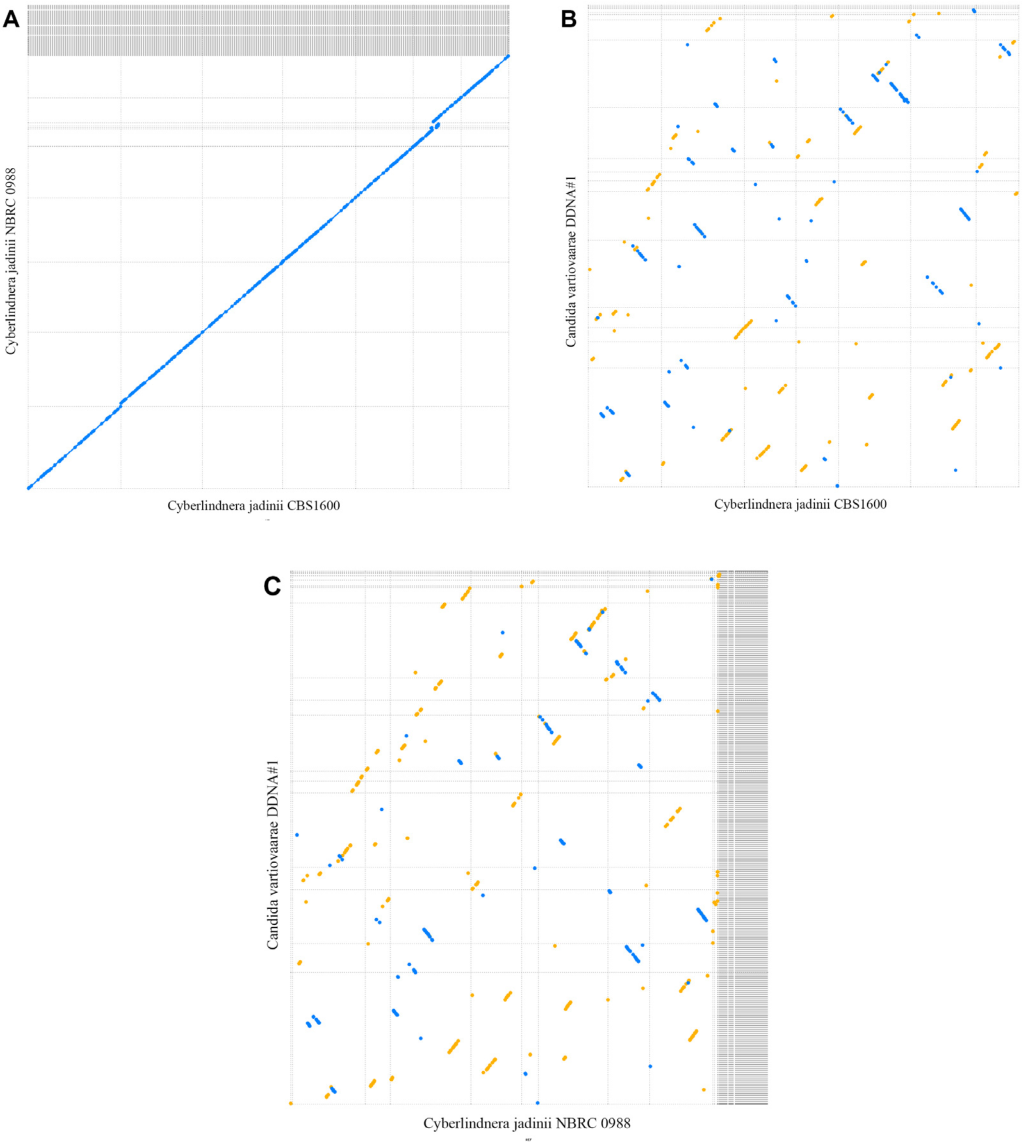

Figure 5 **Full genome comparisons between different yeast species.**

Dashed lines indicate contigs (start and stop positions) and the area between dashed lines indicates the contig size. Blue and yellow dots are hits in reverse and forward orientation, respectively. Diagonal lines indicate sequence and synteny conservation across species. (**A**) Comparison between NBRC 0988 (vertical axis) and Cyberlindnera jadinii strains CBS1600 (horizontal axis) with 8 kbp as minimal hot length. (**B**) Comparison between *Candida vartiovaarae* isolate DDNA#1 (vertical axis) and *Cyberlindnera jadinii* strain CBS1600 (horizontal axis) with 100 bp as minimal hit length. (**C**) Comparison between *Candida vartiovaarae* isolate DDNA#1 (vertical axis) and *Cyberlindnera jadinii* strain NBRC 0988 (horizontal axis) with 100 bp as minimal hit length.

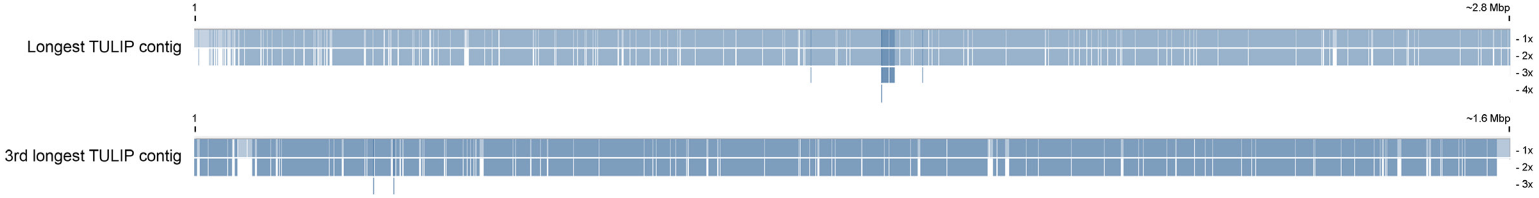

Figure 6 **Tablet visualization of Spades hybrid contigs aligned to TULIP contigs.**
The Spades hybrid contigs aligned against longest TULIP contig (~2.8 Mbp) and the third longest TULIP contig (~1.6 Mbp). White horizontal lines indicate coverage boundaries and show that most regions on the TULIP contigs are covered twice. Alignment gaps come from heavily fragmented Spades hybrid contigs that are aligned on contiguous TULIP contigs. Visualization is based on coverage overview settings in Tablet.

## Author contributions

HPS conceived the study. PJP, HPS, HJJ, and RPD designed the experiments. HJJ, RJLFL, PvH, TO, and SS performed the experiments. HJJ, ML, and CVH contributed to the data analysis. HJJ, RPD, and HPS prepared the first draft of the manuscript.  ML performed additional revision analysis and finalized the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

## Competing interests

HJJ and CVH are members of the Nanopore Community, and have previously received flow cells free of charge, as well as travel expense reimbursements from Oxford Nanopore Technologies.

## Grant information

The author(s) declared that no grants were involved in supporting this work.

## References

1. Zhang GC, Liu JJ, Kong II, et al.: Combining C6 and C5 sugar metabolism for enhancing microbial bioconversion. Curr Opin Chem Biol. 2015; 29: 49–57.

2. Sànchez Nogué V, Karhumaa K: Xylose fermentation as a challenge for commercialization of lignocellulosic fuels and chemicals. Biotechnol Lett. 2015; 37(4): 761–772.

3. Zha Y, Hossain AH, Tobola F, et al.: Pichia anomala 29X: a resistant strain for lignocellulosic biomass hydrolysate fermentation. FEMS Yeast Res. 2013; 13(7): 609–617.

4. Harner NK, Wen X, Bajwa PK, et al.: Genetic improvement of native xylose-fermenting yeasts for ethanol production. J Ind Microbiol Biotechnol. 2015; 42(1): 1–20.

5. Simpson JT, Pop M: The theory and practice of genome sequence assembly. Annu Rev Genomics Hum Genet. 2015; 16: 153–172.

6. Koren S, Phillippy AM: One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015; 23: 110–120.

7. Urban JM, Bliss J, Lawrence CE, et al.: Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. BioRxiv. 2015.

8. Berlin K, Koren S, Chin CS, et al.: Assembling large genomes with singlemolecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015; 33(6): 623–630.

9. Chakraborty M, Baldwin-Brown JG, Long AD, et al.: Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016; 44(19): e147.

10. Marçais G, Kingsford CA: A Fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27(6): 764–770.

11. Vurture WG, Sedlazeck FJ, Nattestad M, et al.: GenomeScope: fast referencefree genome profiling from short reads. Bioinformatics. 2017; 33(14): 2202–2204.

12. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

13. Magoč T, Salzberg SL: FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27(21): 2957–2963.

14. Bankevich A, Nurk S, Antipov D, et al.: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19(5): 455–477.

15. Koren S, Walenz BP, Berlin K, et al.: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. BioRxiv. 2016.

16. Li H: Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics. 2016; 32(14): 2103–2110.

17. Jansen HJ, Liem M, Jong-Raadsen SA, et al.: Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. Sci Rep. 2017; 7(1): 7213.

18. Ruan J: Ultra-fast *de novo* assembler using long noisy reads. 2016 (Januari 2018, date last accessed).

19. Sović I, et al.: (Januari 2018, date last accessed).

20. Walker BJ, Abeel T, Shea T, et al.: Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9(11): e112963.

21. Waterhouse RM, Seppey M, Simão FA, et al.: BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017; 35(3): 543–548.

22. Chen B, Huang X, Zheng JW, et al.: Candida mengyuniae sp. nov., a metsulfuronmethyl-resistant yeast. Int J Syst Evol Microbiol. 2009; 59(Pt 5): 1237–1241.

23. Kurtz S, Phillippy A, Delcher AL, et al.: Versatile and open software for comparing large genomes. Genome Biol. 2004; 5(2): R12.

24. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Oxford University Press, 2013; 1–3.

25. Milne I, Stephen G, Bayer M, et al.: Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform. 2013; 14(2): 193–202.

26. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4): 357–9.

27. Li H: Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics. 2016; 32(14): 2103–10, arXiv: 1512.01801.

28. Xu J: Fungal DNA barcoding. Genome. 2016; 59(11): 913–932.

29. Ip CL, Loose M, Tyson JR, et al.: MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved]. F1000Res. 2015; 4: 1075.

30. Jain M, Tyson JR, Loose M, et al.: MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry [version 1; referees: 1 approved, 2 approved with reservations]. F1000Res. 2017; 6: 760.

— Chapter 3

# Genome assembly of the transposon-enriched *Allorhizobium* strain LBA9072

Chapter in preparation for publication

## — Abstract

The assembly of a transposon-enriched in-house *Allorhizobium* strain was improved from 154 contigs to two circular chromosomes and two additional plasmids using Oxford Nanopore Technologies (ONT) long-reads. We have assembled the sequencing data using assemblers Unicycler, Flye and Canu, using hybrid and *de novo* assembly strategies.
Assembly differences are specifically apparent for the Canu assembly, where the large chromosome was still separated into two distinct contigs and unable to circularize a plasmid. Both hybrid and *de novo* assembly results show high sequence similarity compared to a reference, although some misassemblies are found within the Canu assembly. The frequency and location of two prominently present transposons are identified in addition to transposable elements found by the ISfinder tool.
The lack of sequence similarity between the reference and the final assembly around transposon locations suggest that genomic diversification is facilitated by transposons.

## — Introduction

Here we investigate the genomic structure of an in-house *Allorhizobium* strain (LBA9072), recently reclassified and previously considered an *Agrobacterium* strain[1]. Bacteria of the genus *Agrobacterium* are soil-borne plant pathogens that cause crown gall disease and are also used extensively in genetic engineering[2, 3]. Its genome is usually of moderate complexity, consisting of 2 chromosomes (both circular), a tumor-inducing plasmid (pTi, usually ~200 Kb) and a larger 'cryptic plasmid' of unknown function, and sometimes additional plasmids[4–6]. This pathogen can transfer and integrate a part of its genome (the tumor-inducing T-DNA, found on pTi) into a plant host, reprogramming cells to a proliferation state/ phenotype and resulting in plant tumor formation. Gene expression patterns of host plants show different characteristics which depend on bacterium strain, specialization of strains, plant species and infected cell-type[7]. The genomic structure of *Allorhizobium* is comparable however it contains two circular chromosomes. The genomic structure of the strain investigated in this study deviates from most well-studied strains since it contains large numbers of transposable elements that occur at multiple locations throughout the genome. The lengths of those transposons are longer that Illumina sequencing reads, hence preventing whole genome assembly using Illumina data alone. Since the genomic structure of bacterial genomes is highly versatile it is important to investigating the genomic structure of individual strains. In this study we investigate in detail how long-read sequencing data enables the assembly of complex genomic content, potentially resulting in chromosome-scale contigs.

## — Materials and method

### Illumina sequencing and Velvet assembly on Illumina data

We have generated 99-nucleotide paired-end reads on Illumina HiSeq with 150x coverage for our strain LBA9072 and used Velvet (version 1.2.03, k=63) [v1.1] to assemble the genome. Assembly statistics were calculated with custom Perl scripts and the assembly graph was visualized in Cytoscape [v3.4.0].

### Initial nanopore sequencing and data processing

We produced long reads from genomic DNA using Oxford Nanopore Technologies (ONT) R6 and R7 chemistry and aligned the long reads to the contigs exported by Velvet using LAST (version 4.60)[8]. We used simple settings: gap existence and extension penalties, mismatch penalty, and match reward all equal to 1. Additionally, we used the parameter -m 1000 to increase the alignment hit length until the hit occurs no more than a thousand times on the reference. This increases the number of alignments reported by sacrificing the precision, leading to a more reliable contig tiling across reads while allowing some erroneous alignments in the final report.

### ONT sequencing

We have isolated additional genomic DNA from *Agrobacterium* strain LBA9072 strain using QIAGEN gravity-flow columns and produced another sequencing dataset with 400ng high molecular weight gDNA using R9.4 chemistry. We used a Rapid kit library preparation (SQK-RBK004) according to the manufacturer's protocols (Oxford Nanopore Technologies, Oxford, UK) that allows for swift preparation (approximately 10 minutes) and sequenced for 48 hours granting MinKNOW software (v19.06.8) control to the MinION sequencing device.

### Assembly with Unicycler, Flye and Canu

Long-read data used for assembly were filtered on both length and quality.
The Canu assembly was performed with >1,000 bp reads without quality threshold, a minimum overlap of 500 bp and a 1,000 reads target coverage for read correction.
Unicycler and Flye have been restricted to use reads >3,000 bp that surpass the read quality threshold >10 PHRED, on a modest desktop (7 GB RAM and 8 CPU's) running Ubuntu 16.04 LTS. Additionally, we have provided a 5 Mbp genome size estimate. Flye (V2.4.2)[9] was used to perform *de novo* assembly using ONT data under default settings and using a minimal overlap length of 4,000 bp before considering merging contigs together. Unicycler (V0.4.7)[10] first uses Spades[11] to generate a short-read based assembly graph and performs error correction using short read data, then uses long reads to scaffold short-read based contigs.
Here assembly mode 'normal' (default) was used which is a setting that produces a balanced trade-off between genome completeness and assembly correctness. Gaps between contigs from high quality sequences are then filled in with long-reads, error corrected with Racon[12] and polished with Pilon (v1.18)[13]. Similar data filtering settings were provided, using only reads >3,000 bp with qualities >10 PHRED.

## Transposon count and assembly similarity verification using Mauve

Mauve (v2.4.0)[14] was used to perform full genome progressive pairwise alignments to identify similarity among *de novo* assembly results, and between assemblies and reference strain *Agrobacterium vitis* S4. For circular assembly sequences we have reordered base positions and repositioned the cut generated by the individual assemblers to facilitate homology visualization. Additionally, we aligned two transposon sequences to the final Unicycler assembly result to identify the number of occurrences and location of those repeat sequences.

## Insertion element identification with ISfinder software

ISfinder[15] was used to identify insertion sequences in the Unicycler assembly. ISfinder uses BLAST queries against a database of insertion elements to identify the family the insertion element originated from, as well as the covered length and homology identity to the reference. We have restricted identification to a minimum length of 300 bp, equivalent to a short protein of 100 amino acids and exceeding the length of Illumina reads.

## Assembly visualizing

We used the Circos package (v0.69)[16] to visualize results in comparison to our Unicycler assembly results. We have visualized sequence similarities to genes originating from the *Agrobacterium vitis* S4 reference genome. Genes are aligned to the Unicycler contigs, both start and end positions from full and partial alignments are then converted to .bed file format and used as input for the Circos visualization. Similarly, we have generated bed files for locations of insertion elements identified by ISfinder and locations of two target transposon sequences. Sequencing data coverage of both Illumina and ONT sequencing technologies come from alignment files, and finally, we have compared the initial Velvet assembly to the Unicycler contigs. We have used bed files to report start and end position that we have retrieved from Minimap (V2.17-r954-dirty)[17] alignment files.

Illumina sequencing data quality and Velvet assembly

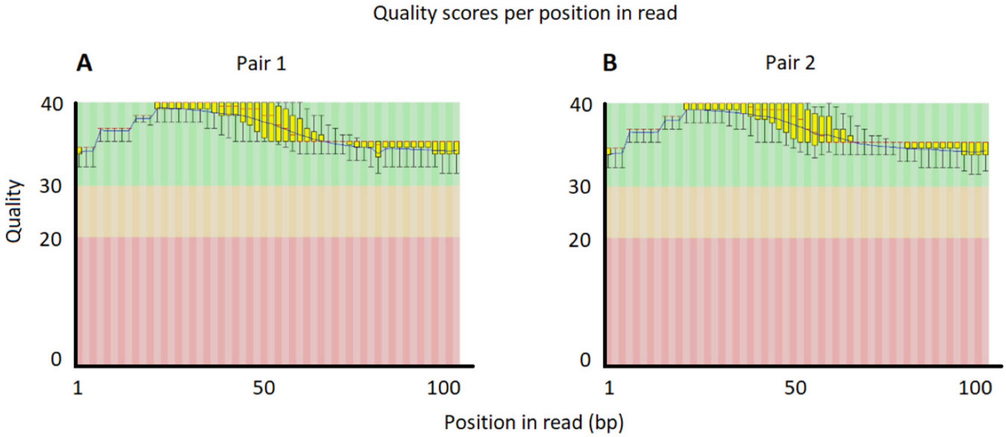

Figure 1 A–B Illumina sequencing data quality control of paired-end sequencing data.

Inspecting high-quality reads reveals sufficient quality for our paired-end dataset (>30 Phred). However, sequences are relatively short (max 100 bp in length), and quality drops are observed at the start and end of reads following the known sequencing characteristics corresponding to Illumina technology sequencing (Figure 1). Those reads are used as input sequences to generate a high-quality whole genome assembly using Velvet and combined with ONT sequencing data for hybrid assembly.

Velvet assembly graph visualisation and transposon connectivity

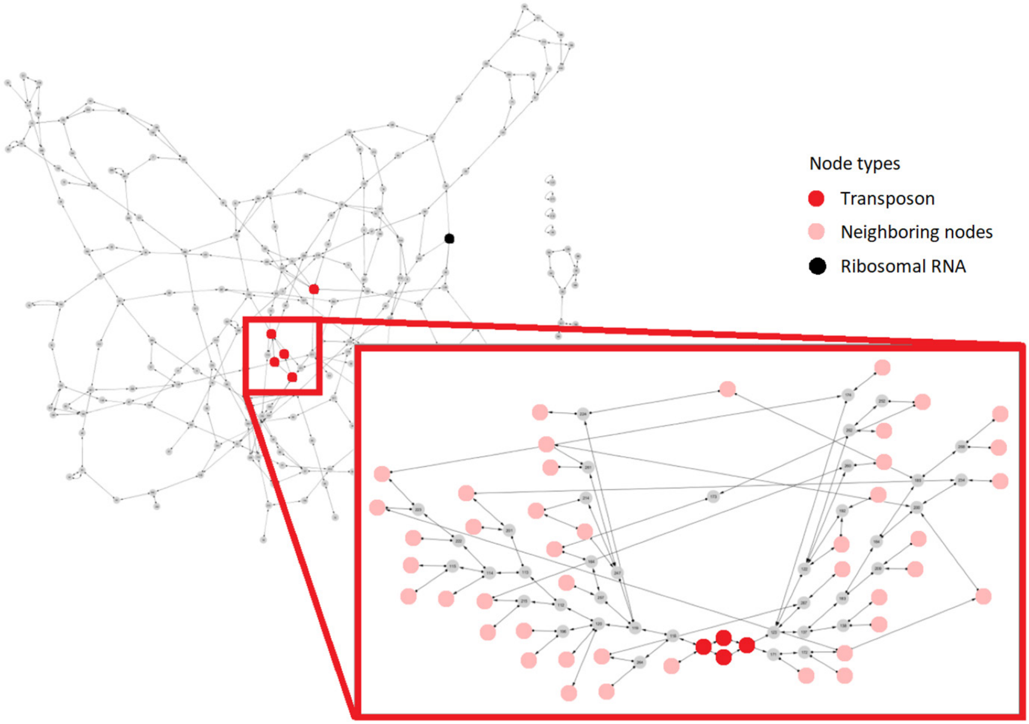

Figure 2 **High-quality whole genome assembly using Velvet.**
Complicated assembly graph due to present transposons indicated by red nodes. Ribosomal RNA contigs are depicted by a black node. The transposon depicted in the red box is split into four nodes due to single nucleotide differences. Contig nodes flanking the transposon are indicated in pink, grey indicates short (k-mer sized) nodes connected to neighbouring contigs.

Figure 2 shows contiguous sequences represented as nodes including the direction in which nodes are connected to their neighboring sequences (input and output directionality), and graph edges represent connections. Nodes that have increased coverage indicate repetitiveness and the coverage allows an estimation on how many times that sequence is observed throughout the complete assembly.

The Illumina-based assembly of strain LBA9072 strain is relatively fragmented, consisting of 154 contigs, with an N50 of 197,590 bp and a total assembly length of 5,872,508 bp (Velvet version 1.2.03, k=63, Table 3). Inspecting the Velvet assembly revealed the genome assembly graph is mainly complicated by the presence of two transposons, one of length 1,285 bp (Figure 2 zoomed-in section four red nodes) and another of length 935 bp (Figure 2 unboxed red node). These repeat contigs are connected to a multiplicity of neighboring nodes. The boxed repeat in red is connected to a total of 43 contigs, of which seven are connected on both left and right side of the repeat (Figure 2 zoomed-in section). Those 50 connections therefore initially suggest that this element is present 25 times throughout the genome. From here on it is referred to as the major transposon. The unboxed red repeat is similarly connected to a set of eight neighboring nodes and suggest the element is present eight times, from here on referred to as the minor transposon.

Sequences of nodes neighboring the major transposon (Figure 2 zoomed-in section pink nodes) are connected to tiny segments (Figure 2 zoomed-in section in gray), typically kmers representing 1-2 bp that do not result in contigs. A detailed view of neighboring nodes and transposon connectivity reveals the difficulty of resolving such context (Figure 2 zoomed-in section). Furthermore, the repeat in red itself is already split into 4 nodes, because of single nucleotide differences. In addition to this complexity, there are several copies of the 6.4 Kbp genes encoding ribosomal RNA (Figure 2 in black). The presence of this ensemble of repeat sequences results in the generation of a complex assembly graph, from which not a single complete plasmid or chromosome can be easily extracted using Illumina data alone.

## Initial nanopore sequencing and data processing

From an experimental Oxford Nanopore Technologies sequencing run we obtained ~13x coverage, 13,158 sequencing reads, with a mean length of 5,611 bp and 12,211 bp N50 length. As the length of these reads often exceeds the lengths of the repeat elements, they could potentially be used to untangle complex contig connections that Illumina data alone cannot resolve (Figure 3 A)

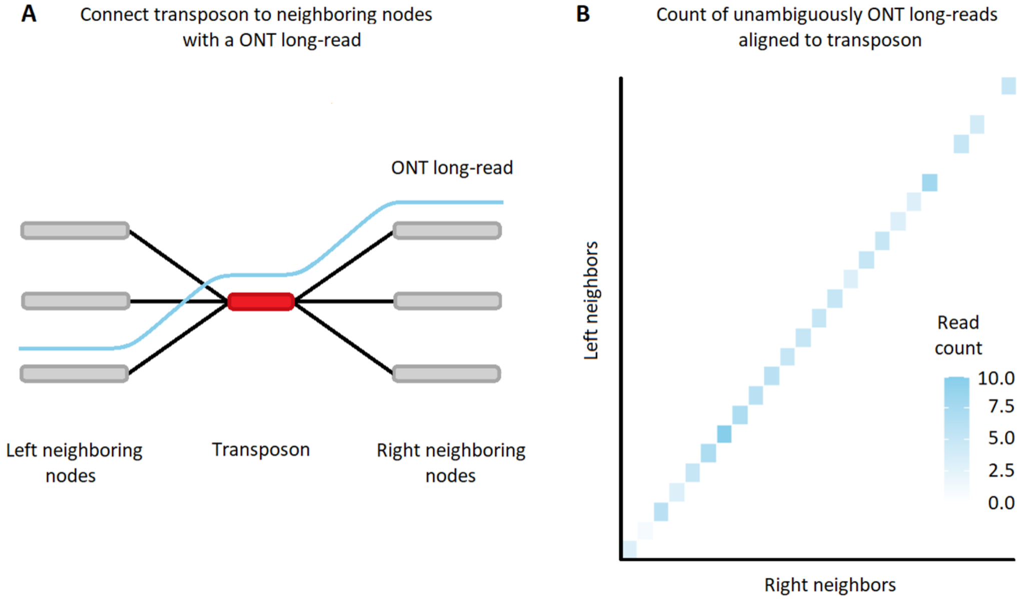

Figure 3 A) schematic representation of a long-read that spans over a repetitive element (**in red**) merging two previously unresolved regions (**in grey**) into a single large contig. **B)** read alignment count that connect left and right neighbors of the major transposon, alignment count ranges from one to ten reads and for two neighbors no alignment was observed.

Since a typical Velvet contig length is much larger compared to the read-length of this sequencing run most alignments are found in the middle of a Velvet contig. For repeat resolution we required a minimum of 2 independent reads that span over a repeat and connect the flanking contigs, in addition to an approximately correct distance between those contigs. Of the 13,158 reads, 483 aligned unambiguously to multiple contigs, and 585 links between contigs could be distilled. From those 585 links a subset is used to connect the major transposon: we analyzed 25 left and right neighboring nodes that are potentially connected to the major transposon (Figure 3 B). Rows represent the 'left' neighbors, columns the 'right'. For almost every neighbor, there exist sufficient unambiguous links (between 3 and 10 reads) to a single other neighbor. One link has low evidence (a single alignment), and for two neighbors no evidence for a connection was observed. In those latter cases, the final placement could not be resolved since neighboring nodes themselves were present twice in the genome.

Using this linkage information, we were able to give our assembly a significant upgrade. However, read-lengths from this sequencing run are insufficient to resolve the 6.4 Kbp ribosomal RNA repeat and these manual curations are very labor intensive. Hence using this low coverage ONT sequencing dataset a complete assembly remained unobtainable.

## ONT sequencing

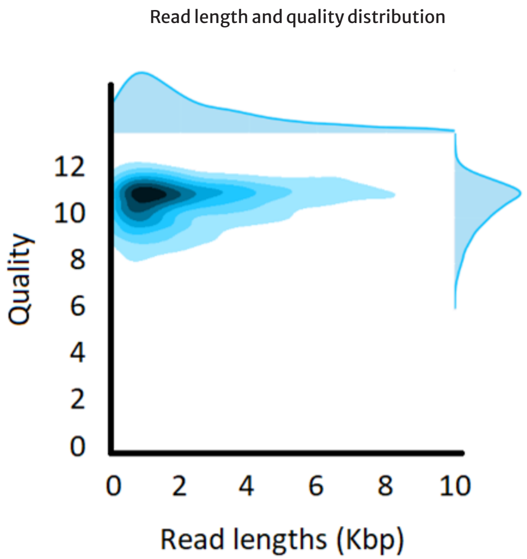

Figure 4 Data quality visualization; quality on Phred scale, where Phred 10 equals 10% error rate.

Table 1 ONT sequencing data statistics

| Statistics | Count |
|---|---|
| Number of reads | 170,955 |
| Number of nucleotides (bp) | 599,185,943 |
| Maximum length (bp) | 63,311 |
| Mean length (bp) | 3,504 |
| Minimum length (bp) | 1 |
| Median length (bp) | 2,121 |

We obtained 600 Mbp in approximately 171,000 reads of sequencing data, which corresponds to approximately 120x coverage of a 5 Mbp genome (Table 1). The average read-quality was better than 10 on the Phred scale (10% error or less) and mean read-lengths were around 3,500 bp (Figure 4). Read-length varies between 1 – 63,311 bp in length. These reads were filtered on quality and length and then used as input sequences for *de novo* assembly and combined with previously described Illumina sequencing data for hybrid assembly.

## Two different assembly strategies

We have tested both long-read-only and hybrid *de novo* assembly strategies on this strain, with Canu and Fly using only ONT reads and Unicycler using both ONT and Illumina data. The resulting assemblies are similar in total genome size, number of contigs, contig lengths and sequence similarity. Using long-read data we have decreased the number of contigs to 4, 5 and 7 contigs for Flye, Unicycler and Canu assembly results, respectively, compared to the Velvet assembly counting 154 contigs (Table 2). Despite different contig number and N50 lengths, total assembly lengths remain similar between all four results. The N50 length is very comparable between Flye and Unicycler, but much lower for Canu. Upon closer inspection of assembled contigs (Figure 5), the main reason for this appears to be that Canu fails to assemble the large single chromosome into a circular contig, but instead outputs two linear contigs. In addition, Canu reports a set of smaller contigs (28,042 and 17,253 bp) that have no clear counterpart in either the Flye or the Unicycler results (Figure 6 B triple asterisk). Finally, since only Unicycler uses sequencing data from two distinct platforms, only Unicycler was able to reconstruct the 5,386 bp phiX174 Illumina spike-in viral genome.

Assembly visualization of two different assembly strategies

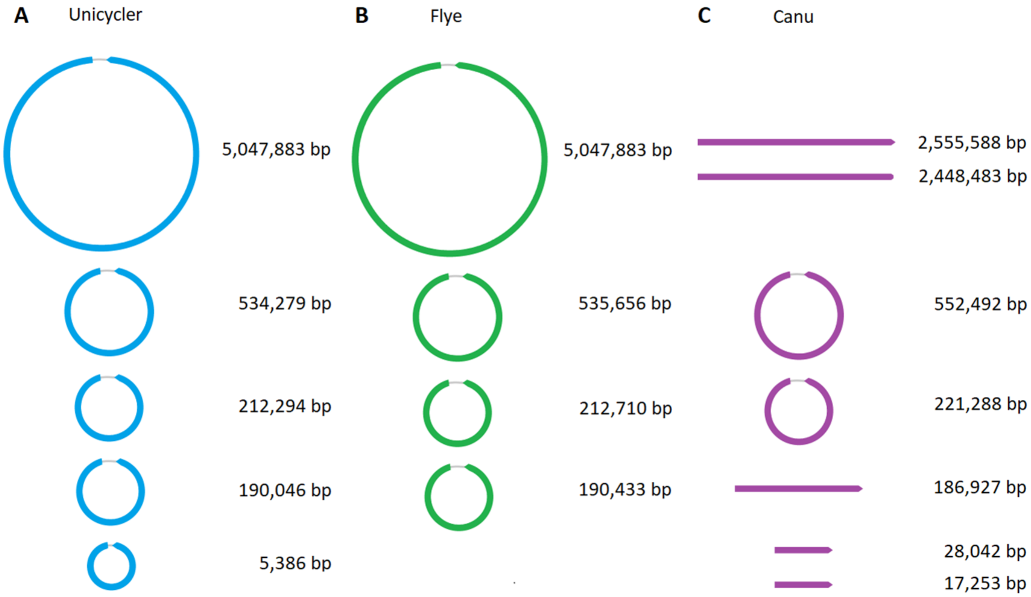

Figure 5 **A)** Unicycler assembly yields five circular contigs, one large chromosome and four additional plasmids. **B)** Flye assembly results in four circular contigs, one large chromosome and 3 additional plasmids similar in length compared to the Unicycler contigs. **C)** Canu outputs 7 contigs; Canu failed to circularize the large chromosome introducing at least two cuts that generate two individual linear contigs. Similarly, a 3 Kbp region is absent from the 190 Kbp plasmid, hence it remains linear.

Table 2 **Assembly statistics**

| Statistics | Velvet | Unicycler | Flye | Canu |
|---|---|---|---|---|
| *Number of contigs* | 154 | 5 | 4 | 7 |
| *Total assembly length* | 5,872,508 | 5,989,889 | 5,999,624 | 6,030,074 |
| *Contig N50* | 197,590 | 5,047,884 | 5,060,825 | 2,468,483 |

### Verifying assembly sequence similarity using Mauve and counting transposon copies

We subsequently investigated the structural quality of the assembled contigs by assessing their similarity to an *Allorhizobium* vitis strain with a fully assembled genome (refered to as S4)[18]. Comparing the assembly results to the S4 reference strain shows overall high sequence and structural similarity between all three assembly results. A striking feature is that the assembled chromosomes have high similarity to the two chromosomes of the reference, and that the largest differences are observed across the plasmids. This suggests that most essential genes are conserved on the chromosome and that the bacterium harbors 'accessory' genes using plasmid sequences. Despite the well-conserved structure between assembly and reference some genomic rearrangements are present. Due to those rearrangements and two additional cuts generated by the Canu assembler the alignment becomes rather complicated to interpret. A single large circular chromosome is presented for both Unicycler and Flye assembly results, whereas Canu results two linear contigs. Some genomic rearrangements are confirmed between all three assembly results. Among others, around locus 1.6 Mbp on our reference A. vitis S4 we observe a small region (Figure 6 A single asterisk) that is reversed and joined on the Unicycler, Flye and Canu assemblies (reverse sequences are depicted on the bottom side of the grey horizontal axis). On the Canu assembly this region is found at 4.7 Mbp since the two largest contigs are linear, hence sequence regions cannot be repositioned to facilitate the visualization (Figure 6 A). Another rearrangement is observed around locus 3.8 Mbp on the reference (Figure 6 A double asterisk) and corresponds to the same green region around 3.0 Mbp on the Unicycler and Flye assembly, and around 3.7 Mbp on the Canu assembly.

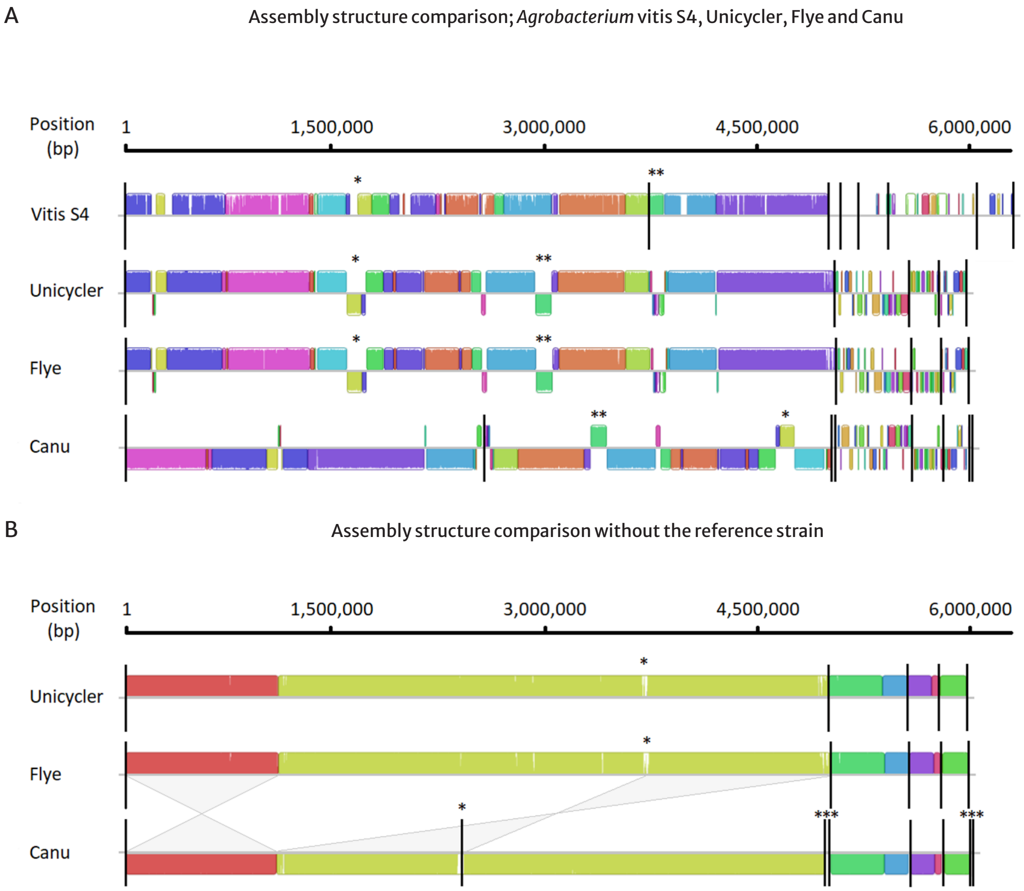

Figure 6 **A)** comparing the reference strain A. vitis S4 to all three assembly results. Coloring indicates similarity between segments, block heights indicate similarity between sequences. Black lines reveal contig ends and the grey horizontal line indicated directionality relative to the reference (forward orientation above the grey axis and reverse on the bottom half). **B)** The same assembly comparison without the S4 reference sequence to facilitate visualization.

Once we remove the reference strain and reorder the circular chromosomes from Unicycler and Flye, the visualization provides a much clearer impression on the quality of the assembly results. All three assemblies show high similarity to each other (Figure 6 colored block heights indicate sequence similarity). The two large linear Canu contigs are particularly misleading since a typical *Allorhizobium* genome comprises two larger chromosomes. Around 3.7 Mbp there exists a small region with low sequence identity between Flye, Unicycler and Canu assembly results (Figure 6 B asterisk), interestingly this is located on one of the cuts introduced by Canu. Finally, the Canu assembly is structurally similar to both Flye and Unicycler results, however, Canu outputs its final contigs in reverse order. The start and end regions of Flye and Unicycler contigs are reversed and merged on the Canu assembly (Figure 6 B in red and green shaded areas, respectively). Furthermore, a low similarity region observed at the same position for Unicycler and Flye assemblies, and around the boundary of the first Canu contig (Figure 6 B single Asterix) in addition to some very small contigs (Figure 6 B triple Asterix).

From the Velvet assembly we retrieved the sequences of the two most prominent repetitive regions. Sequencing data coverage initially suggested those regions were present 25 and 8 times throughout the whole genome. By aligning those sequences to the Unicycler contigs we were able to verify that there are 25 and 9 copies of those regions respectively (Figure 7), consistent with the initial estimates based on the Velvet graph (Figure 2).

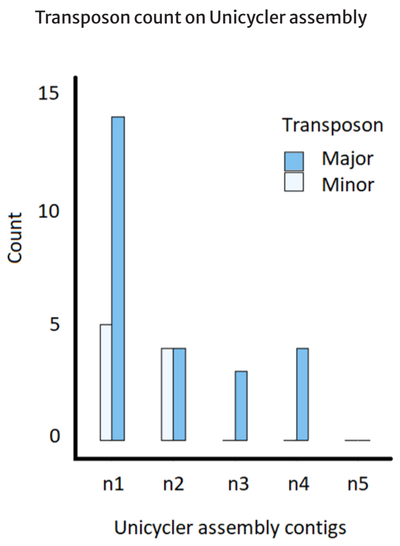

Figure 7 **transposon count on Unicycler assembly contigs.** The two transposon sequences retrieved from the Velvet assembly aligned to the Unicycler contigs. We found 25 copies of the major transposon and 9 copies of the minor transposon.

## Canu discontinuities based on ambiguous alignments

When aligning the two longest Canu contigs to the longest Unicycler contig and zooming in to the regions where the Canu contigs are disjoint we find either a small gap in between or a small overlap of the two Canu contigs. The top track indicates the locations of the Canu cuts on the Unicycler contig, followed by the position track in Mbp, the alignment position of gap and overlap region of Canu contigs and ambiguous read alignments on those regions (Figure 8 blue for ONT data and red for Illumina data). We found ambiguous alignments either in proximity (Figure 8 A) or directly on the cut location (Figure 8 B), for the gap and overlap region, respectively.
This means reads align to the visualized location but also elsewhere in the genome, highlighting an unresolvable decision for assemblers based only on these reads. Despite a generous 80 times coverage of unambiguously aligned reads on those locations (data not shown) Canu is unable to merge the two contigs. A potential explanation for breaking up a contig could be the presence of insertion elements. Those elements have a repetitive nature and could cause reads to align ambiguously and in turn complicate decision-making processes that eventually lead to the introduction of a cut. Insertion elements, among which the major transposon, are situated in proximity to the two loci where a cut is introduced (Figure 9 around 1,330 and 3,800 Mbp).

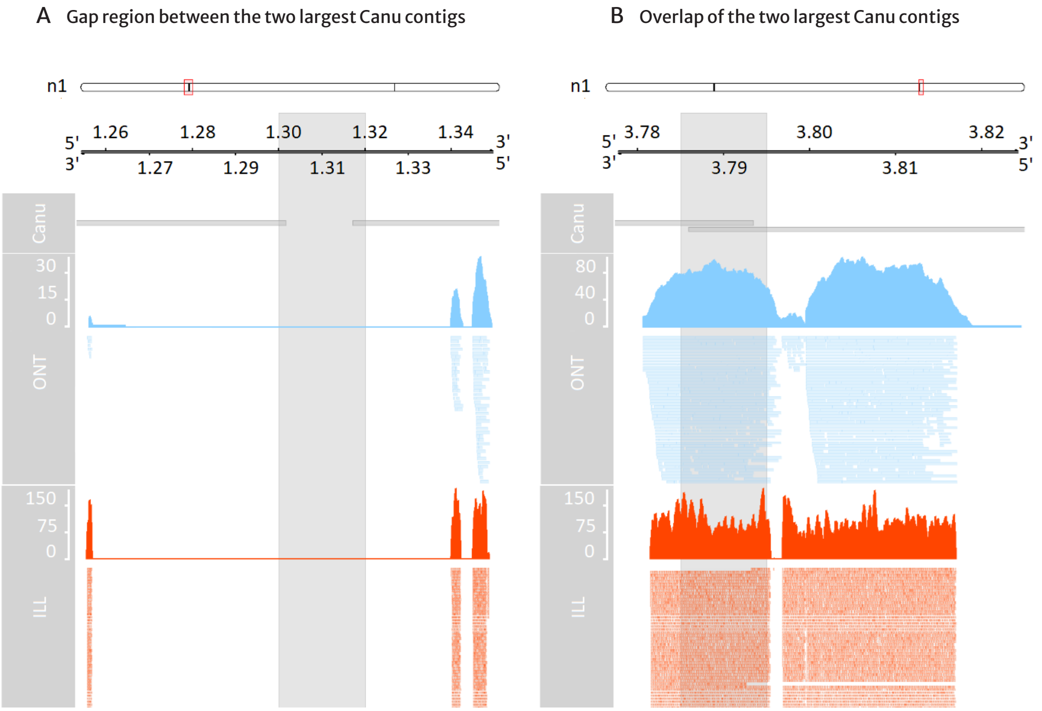

Figure 8 A) Two largest linear Canu contigs aligned to the largest circular Unicyler contig indicated at the top (n1). A small gap in between the two linear Canu contigs is present around 1.3 Mbp. Ambiguous alignments are represented in blue and red for ONT and illumina sequencing data, respectively. Some ambiguities are observed in proximity to the gap region.
B) Around 3.79 Mbp an overlapping region is present near the Canu cut, ambiguous alignments from both ONT and Illumina data, across and around the overlap, offer an explanation why Canu struggles to link those two contigs into a single sequence.

### Unicycler assembly visualization and Velvet assembly comparison

Here we depict an overview of the Unicycler assembly contigs compared to our previous Velvet assembly results. From outwards to inwards (Figure 9), the position track in Kbp, followed by gene homology to the reference strain S4 (black). From the 5,433 genes provided from the reference strain we were able to align 4,400 to the Unicycler assembly. 3,993 genes on the n1 contig, 216 on the n2 contig, 91 on the n3 contig and 99 on the n4 contig. The largest contigs has a higher gene homology compared to the plasmid sequences, interestingly the surroundings of a few insertion sequences show a clear lack of gene similarity (e.g., around 190 Kbp, 400 Kbp, 2,500 Kbp, 3,800 Kbp). Interestingly, three of those loci are identified as insertion elements that are not classified as either the major or minor transposon. The absent gene similarity around those locations suggests that genomic diversification is facilitated by, and therefore found around, repetitive regions. Most insertion elements are observed inside the smaller contigs n2, n3 and n4.

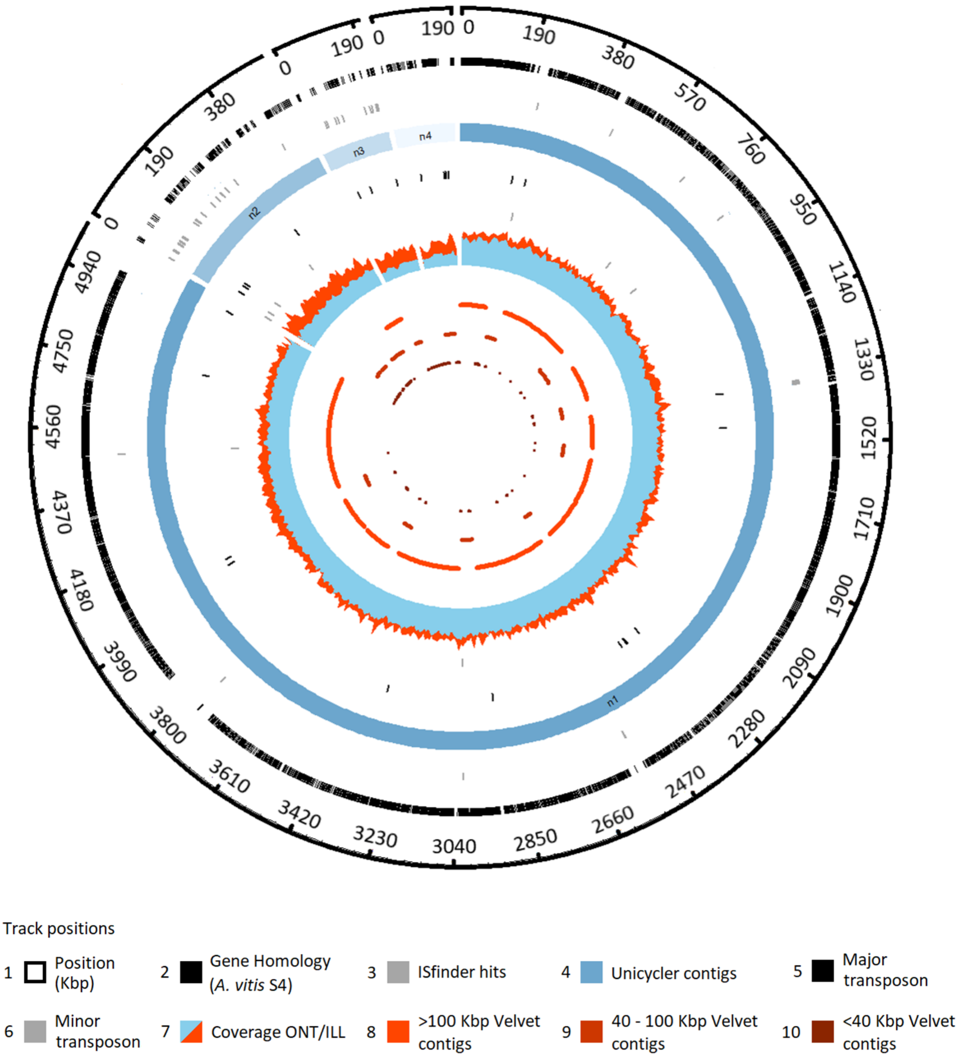

Figure 9

**Track 1;** Base position (in Kbp).

**Track 2;** Shows the Unicycler homology to genes that originate from our reference strain A. vitis.

**Track 3;** Locations of insertion elements identified by ISfinder on the Unicycler contigs.

**Track 4;** Unicycler contigs (excluding n5 that is too small to visualize).

**Track 5;** Locations of the major transposon on the Unicycler contigs.

**Track 6;** Locations of the minor transposon on the Unicycler contigs.

**Track 7;** sequencing data coverage (ONT data in blue overlapping Illumina data in red).

**Track 8;** Velvet contigs >100 Kbp.

**Track 9;** Velvet contigs between 40 and 100 Kbp.

**Track 10;** Velvet contigs smaller then 40 Kbp.

The third track shows insertion sequences as found by the tool ISfinder. ISfinder identifies several different insertion sequences, the major and minor transposon are classified as insertion sequences from the IS5 family. However, it only reports a subset of the major and minor transposon locations compared to aligning the major and minor transposon sequences to the Unicycler contigs using Minimap2 (track five and six). Interestingly, BLAST results of the two transposon sequences from the Velvet assembly originate from the Rhizobium sp. 21/90 tumor inducing plasmid found in a Himalayan blackberry from Oregon USA at locus 104,603 - 105,694 and 188,695 – 189,528[19]. Those hits have >98% identity over the complete transposon region and are both annotated as IS5 family transposases. Alignment to the genes from the S4 reference did not result in significant hits, unless the similarity threshold was relaxed considerably (only 82% of the 935 bp was covered with <75% identity). The fourth track indicates the Unicycler contigs ordered from large to small (indicated by a blue shade from dark to light). The fifth and the sixth track reveal the locations of the major (black) and minor (grey) transposon copies that are aligned to the Unicycler contigs. Track seven is an overlay visualization of Illumina (red) and ONT (blue) data coverage. Interestingly, a large difference in coverage between ONT and Illumina data is observed for plasmid sequences, even though we have not performed a read-length selection. Finally, the eighth to tenth track depict the Velvet assembly results ordered from large to small contigs. Contigs >100 Kbp are indicated in bright red, followed by contigs between 100 Kbp and 40 Kbp and finally in dark red contigs <40 Kbp. The >100 Kbp track indicates the large chromosome is nearly complete, only a few gaps are introduced based on Illumina data alone. However, due to the repetitive structure some noise remains observed in the final track (small contigs overlap with the larger ones). High fragmentation is primarily observed for the plasmids, were some contigs range between 100 and 40 Kbp and many <40 Kbp contigs are observed. This represents low sequence complexity making it more difficult to assemble plasmids accurately.

## References

1. Kuzmanović, N., Biondi, E., Overmann, J. et al. Genomic analysis provides novel insights into diversification and taxonomy of *Allorhizobium* vitis (i.e. *Agrobacterium* vitis). BMC Genomics 23, 462 (2022). https://doi.org/10.1186/s12864-022-08662-x

2. Anu Kalia, Nanotechnology in Bioengineering: Transmogrifying Plant Biotechnology, Omics Technologies and Bio-Engineering Volume 2: Towards Improving Quality of Life 2018, Pages 211-229 | https://doi.org/10.1016/B978-0-12-815870-8.00012-7

3. Indra A. Padikasan et al., Agricultural Biotechnology: Engineering Plants for Improved Productivity and Quality | https://doi.org/10.1016/B978-0-12-815870-8.00006-1

4. Eugene W. Nester, *Agrobacterium*: nature's genetic engineer, Front. Plant Sci., 06 January 2015 | https://doi.org/10.3389/fpls.2014.00730

5. Gustavo A. de la Riva, *Agrobacterium* tumefaciens: a natural tool for plant transformation, Electron. J. Biotechnol. v.1 n.3 Valparaíso dic. 1998 | http://dx.doi.org/10.4067/S0717-34581998000300002

6. J. S. Robalino-Espinosa, Segregation of four *Agrobacterium* tumefaciens replicons during polar growth: PopZ and PodJ control segregation of essential replicons, PNAS October 20, 2020 | https://doi.org/10.1073/pnas.2014371117

7. Jochen Gohlke et. Al., Plant responses to *Agrobacterium* tumefaciens and crown gall development, Front. Plant Sci., 23 April 2014 | https://doi.org/10.3389/fpls.2014.00155

8. Szymon M. Kiełbasa, et al., Adaptive seeds tame genomic sequence comparison | Published in Advance January 5, 2011, doi:10.1101/gr.113985.110Genome Res. 2011. 21: 487-49

9. Kolmogorov M., Yuan J., Lin Y. and Pevzner PA. (2019). Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37(5), 540-546.

10. Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595

11. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. pmid:22506599

12. Robert Vaser et al., Fast and accurate *de novo* genome assembly from long uncorrected reads | Genome Res. 2017 May; 27(5): 737–746. doi: 10.1101/gr.214270.116

13. Walker BJ, Abeel T, Shea T, et al.: Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement | PLoS One. 2014;9(11):e112963. 10.1371/journal.pone.0112963

14. Aaron C.E. Darling et al., Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements | Genome Res. 2004 Jul; 14(7): 1394–1403. doi: 10.1101/gr.2289704

15. Siguier P. et al. (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34: D32-D36 | doi: 10.1093/nar/gkj014.

16. Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645

17. Heng Li, Minimap2: pairwise alignment for nucleotide sequences | Bioinformatics, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191

18. Steven C. Slater et al., Genome Sequences of Three *Agrobacterium* Biovars Help Elucidate the Evolution of Multichromosome Genomes in Bacteria | DOI: https://doi.org/10.1128/JB.01779-08

19. Alexandra J Weisberg et al., Diversification of plasmids in a genus of pathogenic and nitrogen-fixing bacteria | Philos Trans R Soc Lond B Biol Sci. 2022 Jan 17;377(1842):20200466. doi: 10.1098/rstb.2020.0466. Epub 2021 Nov 29.

— Chapter 4

# Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads

Scientific Reports, 7: 7213 , 2017

Hans J. Jansen[1], Michael Liem[2], Susanne A. Jong-Raadsen[1], Sylvie Dufour[3], Finn-Arne Weltzien [4], William Swinkels[5], Alex Koelewijn[5], Arjan P. Palstra[6], Bernd Pelster[7], Herman P. Spaink[2], Guido E. van den Thillart[1], Ron P. Dirks[1] & Christiaan V. Henkel [2,8,9]

[1] ZF-screens B.V., Leiden, The Netherlands.

[2] Institute of Biology, Leiden University, Leiden, The Netherlands.

[3] Muséum National d'Histoire Naturelle, Sorbonne Universités, Research Unit BOREA, Biology of Aquatic Organisms and Ecosystems, CNRS, IRD, UCN, UA, Paris, France.

[4] Norwegian University of Life Sciences, Faculty of Veterinary Medicine, Department of Basic Science and Aquatic Medicine, Oslo, Norway.

[5] DUPAN Foundation, Wageningen, The Netherlands.

[6] Animal Breeding and Genomics Centre, Wageningen Livestock Research, Wageningen University & Research, Wageningen, The Netherlands.

[7] Institute of Zoology and Center for Molecular Biosciences, University of Innsbruck, Innsbruck, Austria.

[8] University of Applied Sciences Leiden, Leiden, The Netherlands.

[9] Generade Centre of Expertise in Genomics, Leiden, The Netherlands.

Correspondence and requests for materials should be addressed to C.V.H. (email: c.v.henkel@biology.leidenuniv.nl)

## — Abstract

We have sequenced the genome of the endangered European eel using the MinION by Oxford Nanopore, and assembled these data using a novel algorithm specifically designed for large eukaryotic genomes. For this 860 Mbp genome, the entire computational process takes two days on a single CPU. The resulting genome assembly significantly improves on a previous draft based on short reads only, both in terms of contiguity (N50 1.2 Mbp) and structural quality. This combination of affordable nanopore sequencing and light weight assembly promises to make high-quality genomic resources accessible for many non-model plants and animals.

## — Introduction

Just ten years ago, having one's genome sequenced was the privilege of a handful of humans and model organisms. Spectacular improvements in high-throughput technology have since made personal genome sequencing a reality and prokaryotic genome sequencing routine. In addition, sequencing the larger genomes of non-model eukaryotes has opened up a wealth of information for plant and animal breeding, conservation, and fundamental research.

As an example, we and others[1–3] have previously established genomic resources for the European eel (*Anguilla anguilla*), an iconic yet endangered fish species that remains resistant to efficient farming in aquaculture[4, 5]. A draft genome[2], several transcriptomes[1, 3–10], and reduced representation genome sequencing[11] have already shed light on its evolution and developmental biology[2, 12, 13], endocrinological control of maturation[7, 9], metabolism[14], disease mechanisms[10], and population structure[15, 16], thereby supporting both breeding and conservation efforts. However, compared to established model organisms, funds for eel genomics are naturally limited, and consequently the quality of current genome assemblies of *Anguilla* species is modest at best by today's standards (Table 1).

The recent availability of affordable long-read sequencing technology[17, 18] by Oxford Nanopore Technologies (ONT) presents excellent opportunities for generating high-quality genome assemblies for any organism[19].
Flow cells for the miniature MinION sequencing device employ a maximum of 512 nanopores concurrently for reading single-stranded DNA at up to 450 nucleotides per second, resulting in several gigabases of sequence during a two day run. As the technology does not rely on PCR or discrete strand synthesis events, DNA fragments can be of arbitrarily long length.
The single-molecule reads are of increasingly good quality, with a sequence identity of ~75% for the older R7.3 chemistry[17], to ~89% for the newer R9 chemistry (MinION Analysis and Reference Consortium, in preparation).
Optionally, DNA can be read twice (along both strands) to yield a consensus '2D' read of higher accuracy (up to ~94% for R9).

Long-read sequencing technology is also offered by Pacific Biosciences (PacBio). This platform employs advanced optics to detect a polymerase operating on single DNA molecules, and has been commercially available since 2011. Both long-read technologies deliver roughly comparable quality and data volumes. PacBio sequencing has the advantages of an established, stable platform (which includes bioinformatics), as well as less bias in the error profile.

Table 1 Previous genome assemblies of *Anguilla* species.
*Not all contigs obtained by *de novo* assembly were used in scaffold construction.

| Species | Reference | NCBI WGS reference | Assembly methods | Contigs/ scaffolds sum | Contig/scaffold N50 | Scaffold gaps |
|---|---|---|---|---|---|---|
| *A. anguilla* | 2 | AZBK01 | CLC bio + SSPACE | 969/923 Mbp* | 1.7/77.6 kbp | 134 Mbp |
| *A. japonica* | 34 | AVPY01 | CLC bio + SSPACE | 1.13/1.15 Gbp* | 3.3/52.8 kbp | 127 Mbp |
| *A. rostrata* | 37 | LTYT01 | Ray + SSPACE | 1.19/1.41 Gbp | 7.4/86.6 kbp | 223 Mbp |

Table 2 *Anguilla* genome size predictions.
*Ranges are the minimum and maximum values reported for three model fits at different k-mer lengths.
Apparent repetitive sequence decreases with k-mer length, and heterozygosity increases with k-mer length.
**For *A. japonica*, the model did not converge in most cases, presumably because of low coverage. These results are for k = 19.

| Species | Haploid genome size* | Repetitive fraction* | Heterozygous fraction* |
|---|---|---|---|
| *A. anguilla* | 854.0–866.5 Mbp | 15.5–20.0% | 1.48–1.59% |
| *A. japonica*** | 1.022 Gbp | 38.7% | 2.74% |
| *A. rostrata* | 799.0–813.0 Mbp | 12.2–16.9% | 1.50–1.60% |

Advantages of ONT include the much lower equipment cost, and currently rapidly improving quality, read length and throughput. Comprehensive comparisons of both technologies are scarce[20].

In contrast to short reads, long reads offer the possibility to span repetitive or otherwise difficult regions in the genome, resulting in strongly reduced fragmentation of the assemblies. This potential advantage does require the deployment of dedicated genome assembly algorithms that are aware of long-read characteristics. In addition, as single-molecule long-read technologies (by both PacBio and ONT) do suffer from reduced sequence identity, this likewise needs to be addressed by post-sequencing bioinformatics[21–23]. Dealing with these challenges has reinvigorated research into genome assembly methodology, resulting in several novel strategies[24–28].

However, when dealing with large eukaryotic genomes, the computational demands for long-read assembly are often higher than for short reads (using De Bruijn-graphs), even though the raw data are more informative of genome structure. Especially now that sequencing very large plant and animal genomes is finally becoming both technologically feasible and affordable, the computational costs may turn out to be prohibitive.
For example, using the state-of-the-art Canu assembly software[25], assembling a human genome from long reads takes tens of thousands of CPU hours, or several days on a computer cluster (https://genomeinformatics.github.io/NA12878-nanopore-assembly). As scaling behaviour is approximately quadratic with genome size, assembling a salamander[29] or lungfish[30] genome dozens of gigabases long would require several years on a cluster.

We are currently developing a computational pipeline specifically intended for future sequencing of extremely large tulip genomes[31] (up to 35 Gbp). Named TULIP (for *The Uncorrected Long-read Integration Process*), its primary purpose is to split up such large assembly problems into manageable subsets of long reads. Each subset can then be handled by a separate downstream *de novo* assembly process, in theory substituting quadratic scaling with nearly linear behaviour. Here, we use a prototype of this algorithm to assemble a new version of the European eel genome, based on Oxford Nanopore sequencing. The entire computational procedure takes two days on a desktop computer, and yields an assembly that is two orders of magnitude less fragmented than the previous Illumina-based draft.

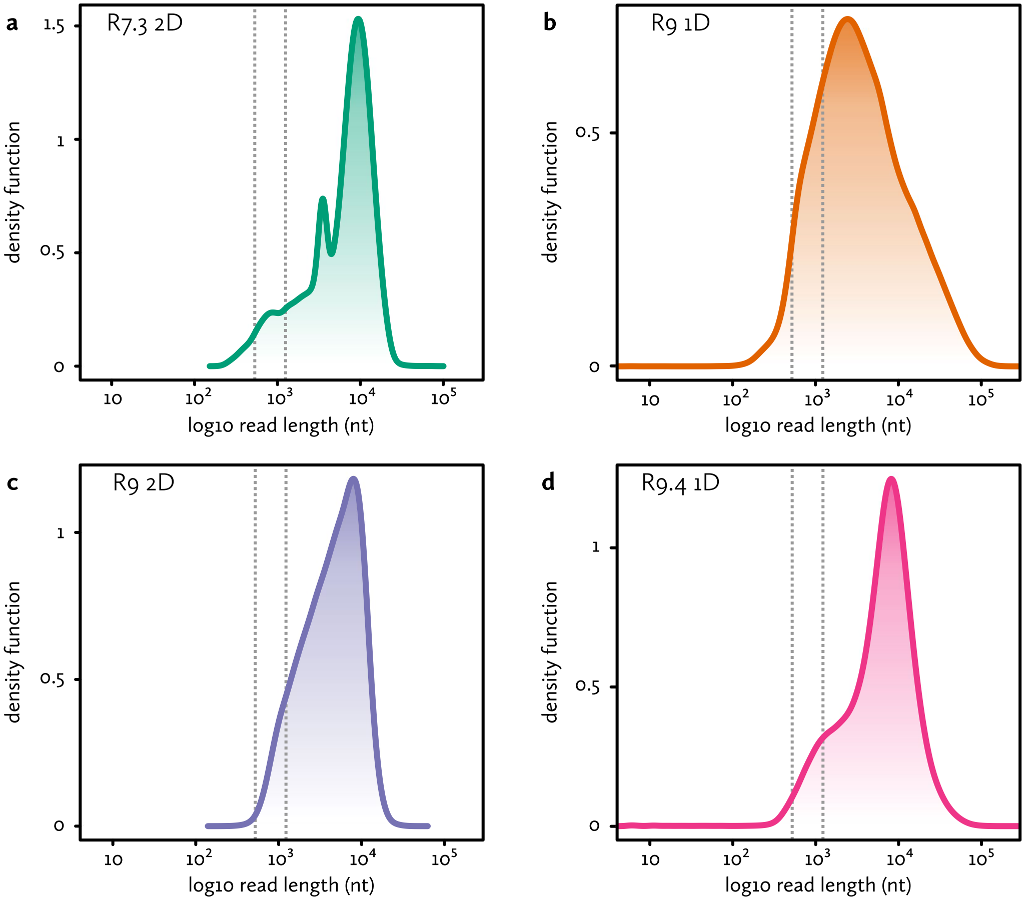

Figure 1 **Nanopore sequencing.**
Shown are the sequenced fragment size distributions for the **(a)** R7.3 chemistry 2D reads, **(b)** R9 chemistry 1D reads, **(c)** R9 chemistry 2D reads and **(d)** R9.4 chemistry 1D reads.
Dotted lines indicate the minimum (542 bp) and typical (1270 bp) read lengths that can be used for linking two seeds in the 0.29× overage 285 bp set. The minimum length is 2 × 285 bp with no more than 10% overlap between seeds. The typical length assumes an average of one seed per 985 bp (genome size divided by number of seeds).

## — Results

### Eel genome sizes and previous assemblies

Before launching a genome sequencing effort, an estimate of the size of the genome of interest is needed. For the genus *Anguilla*, several studies have used flow cytometry and other methods to arrive at C-values ranging from 1.01 to 1.67 pg (http://www.genomesize.com), corresponding to haploid genome sizes in the 1–1.6 Gbp range for both *A. anguilla* and *A. rostrata*. We previously estimated a genome size of approximately 1 Gbp for *A. anguilla*, using human cells as a reference[2]. Based on their assembled genomes, *Anguilla* species exhibit a similarly wide range of apparent genome sizes (see Table 1). These draft assemblies are all based on previous-generation short-read technology, and relied on Illumina mate pairs to supply long-range information used in scaffolding. The resulting assemblies remain highly fragmented, with low N50 values even considering the technology used. We therefore examined k-mer profiles in the raw Illumina sequencing data, which can provide an estimate of the length of the haploid genome[32, 33]. Surprisingly, the predicted genome sizes are considerably – but consistently – smaller than previously estimated or assembled (Table 2 and Supplementary Fig. S1). In addition, all three examined genomes contain high levels of heterozygosity.

### Nanopore sequencing

We isolated DNA for long-read sequencing from the blood and liver of a fresh female European eel. Using three different generations of the ONT chemistry for the MinION sequencer, we generated 15.6 Gbp of raw shotgun genome sequencing data (see Fig. 1 and Supplementary Table S1). Assuming an 860 Mbp haploid size, this corresponds to approximately 18-fold coverage of the genome. The bulk of the sequence is in long or very long reads (up to hundreds of thousands of nucleotides), although a fraction is composed of very short reads or artifacts (e.g. 6 bp reads, Fig. 1). We used all raw reads for subsequent genome assembly.

### Assembly strategy

We assembled the long nanopore sequencing reads using a prototype of an assembly strategy we are developing for very large genomes (M. Liem and C. Henkel, in preparation), named TULIP. Briefly, it takes two shortcuts compared to the established hierarchical approach[21, 25].
First of all, like Miniasm[27], TULIP does not correct noisy single-molecule reads prior to assembly. Secondly, it does not perform an all-versus-all alignment of reads, but instead aligns reads to a sparse reference (of 'seed' sequences) that is representative for the genome. The result is a 'seed graph', which can be used to either partition the original long reads into many independent subsets for subsequent *de novo* assembly, or to immediately extract uncorrected scaffold sequences from. Here, we have chosen to use the latter functionality, and employed stand-alone post-assembly consensus applications to correct the resulting scaffolds.

Figure 2a illustrates all the steps we have taken during *de novo* assembly of the European eel genome. We employed previously generated Illumina shotgun sequencing reads as sparse seeds. Using a k-mer counting table, we identified merged read pairs that are suitably unique in the genome. Using strict criteria (see Methods), we could select 5019778 fragments of 270 bp, or 873058 of 285 bp, corresponding to 1.58-fold or 0.29-fold coverage of the genome, respectively. We subsequently used several random subsets of these fragments as a reference to align long nanopore reads against.

Using a custom script, we constructed a graph based on these alignments, in which the seed sequences are nodes, and edges represent long read fragments (Fig. 2b). A connection between two seeds indicates they co-align to a long read, and are therefore presumably located in close proximity in the genome. In theory, perfect alignments of very long reads to unique seeds should be sufficient to organize both sets of data into linear scaffolds.

However, because of the errors still present in long nanopore reads, the alignments are imperfect, with missed seed alignments making up the bulk of ambiguities in the seed graph (i.e. forks and joins in the seed path). Additional uncertainties are introduced by spurious alignments and residual apparently repetitive seeds. The tangles these cause in the graph can be recognized locally, and are removed during a graph simplification stage (Fig. 2c). TULIP will visit every seed that has multiple in- or outgoing connections, and attempt to simplify the local graph topology by removing connections. For example, if a single seeds fails to align to a single nanopore read, this will introduce a 'triangle' in the graph (Fig. 2c, top example), in which the neighbouring seeds now share a direct connection (based on that single read). If the intermediate seed fits between the neighbouring seeds, TULIP will then remove the connection spanning the intermediate seed. If after this stage a seed still has too many connections, it might represent repetitive content and its links are severed altogether (Fig. 2c, second example).

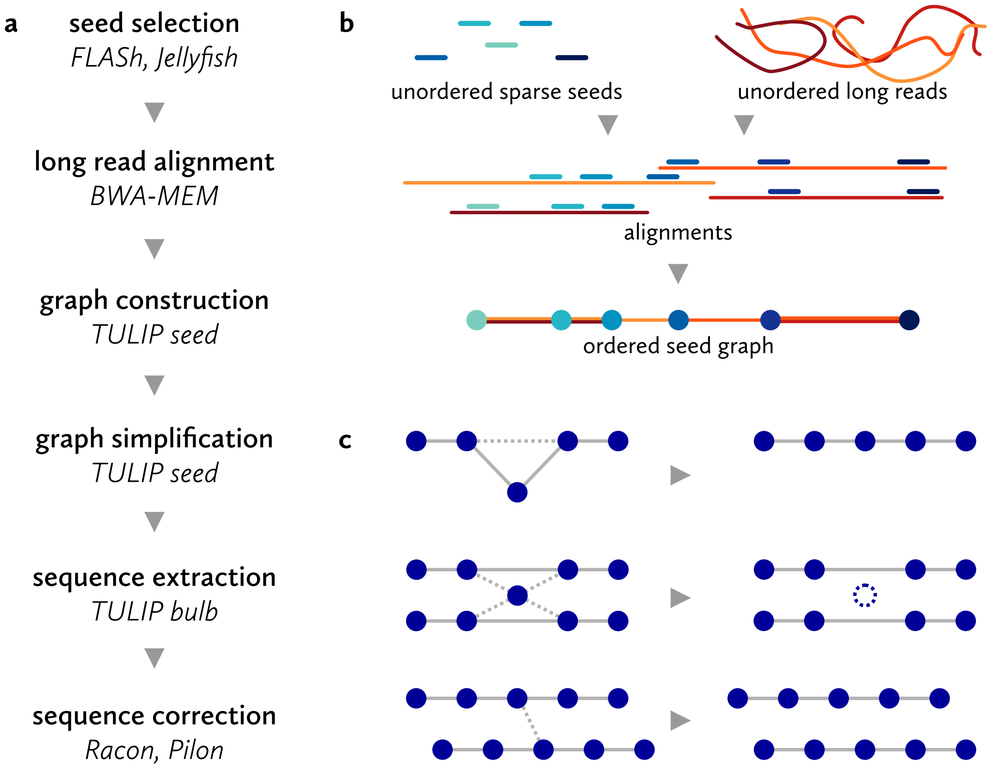

Figure 2 **Assembly strategy.**
**(a)** Stages in the TULIP assembly of the European eel genome.

**(b)** Graph construction based on long read alignments to short seeds. Seeds are included in the graph as nodes if they align adjacent to each other to a long read. The apparent distance between the seeds is included as an edge property, as is the amount of evidence (i.e. number of alignments supporting the connection).

**(c)** The initial seed graph based on alignments contains ambiguities, caused by missed alignments, repetitive seed sequences and spurious alignments. These are removed during the initial layout process, resulting in linear scaffolds. Where possible, these scaffolds are subsequently linked by further unambiguous long-distance co-alignments to long reads.

Finally, unambiguous linear arrangements of seeds can be extracted from the graph. Figure 3 illustrates a small fragment of the actual seed graph, with final linear paths (scaffolds) and removed connections indicated. These ordered seed scaffolds do not yet contain sequence data. These can subsequently be added from the original nanopore reads and alignments, resulting in uncorrected scaffold sequences. The scaffolds are exported  bundled with their constituent nanopore reads, and can be subjected to standard nanopore sequence correction procedures.

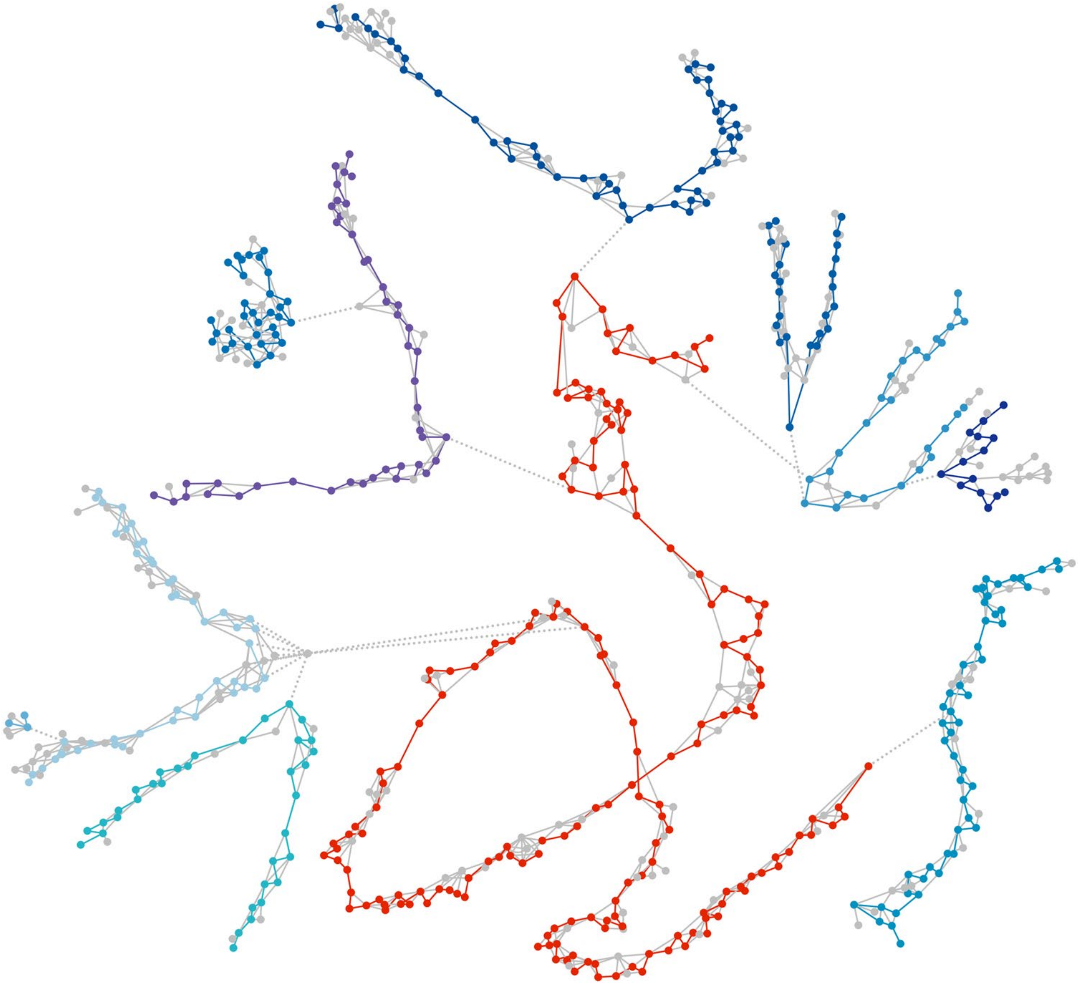

Figure 3 **Graph simplifications.** Scaffolds were extracted from a graph consisting of seed sequences (nodes) linked by nanopore reads (edges). Here, a small final scaffold (number 2231, 252.2 kbp) is shown in red in the context of the initial seed graph (all seeds at a distance of up to ten links from the final scaffold). Fragments of ten other scaffolds (blues) are directly or indirectly connected to scaffold 2231 by a few incorrect links (dotted lines). Seeds and links removed during graph simplification are shown in grey. Scaffolds can be discontinuous in the initial graph, as additional long-distance links are added in a later stage. The graph was visualized using Cytoscape (version 3.4.0).

## Assembly characteristics

We used several combinations of short seed sequences and aligned nanopore reads to optimize the assembly process. In most cases, we did not complete the entire assembly process by adding actual nanopore sequence. Therefore, distances between seeds (and scaffold lengths) are means based on multiple nanopore reads. Adding specific sequence (and subsequently correcting scaffolds) can change these figures slightly. Supplementary Table S2 lists the assembly statistics for these experimental runs.

Both the contiguity and size of the assembly clearly improve upon adding more nanopore data (Fig. 4a,b). This suggests that at 18-fold coverage of this genome, and using the particular blend of data types available here, the assembly process is still limited by the total quantity of long read data.

For the seeds, we investigated the effects of seed length (270 or 285 bp), as well as seed density (fractions and multiples based on the 873058 fragments available at 285 bp). There does not appear to be a clear advantage to choosing either 270 or 285 bp seeds. At identical densities, the two possibilities yield comparable assemblies in terms of size and contiguity.

For seed density, there does appear to be an optimum. As expected, low densities result in fragmentation and incompleteness (Fig. 4c,d). The assemblies with the highest seed density (1.3 or 1.7 million 270 bp sequences) do yield the highest N50 and assembly sum, but also exhibit increased fragmentation compared to lower seed densities. As Fig. 4c shows, the main difference with those assemblies is the appearance of many small scaffolds at high seed numbers. Accidentally, in this case the optimal seed density is around the 'full' set of 873058 fragments, of either 270 or 285 bp. Both also yield an assembly that is close to the estimated genome length. We selected the 285 bp version as a candidate for an updated reference genome for the European eel.

Figure 4 summarizes several characteristics of the candidate assembly (before sequence addition or correction). The length distribution of the 2366 scaffolds (Fig. 4a) shows they range in size between 431 bp and 8.7 Mbp. The lower boundary is expected, as a minimal scaffold has to consist of at least two 285 bp seeds, and the graph construction was executed with parameters allowing limited overlap between seeds. The cumulative scaffold length distributions (Fig. 4c) show that a considerable fraction of the genome is included in large scaffolds, with 232 scaffolds larger than a megabase constituting 56% of the assembly length. Seeds in the final scaffolds are connected by on average 7.4 nanopore read alignments. As can be seen in Fig. 4e, links removed during the graph simplification stage (mostly based on local graph topology only) were predominantly those supported by less evidence.

The final assembly retains 637792 seeds of 285 bp, equivalent to a maximum of 181.8 Mbp of Illumina-derived sequence. If the seed distribution is assumed to be essentially random (with local genomic architecture responsible for exceptions), the initial 873058 seeds should be spaced at a mean interval of 700 bp. As seeds are removed during simplification, larger 'gaps' filled with nanopore-derived sequence should appear. However, as Fig. 4f shows, gap lengths are heavily biased towards low and negative lengths (i.e. overlapping seeds). In this case, this could be an artifact of the very stringent seed selection procedure.

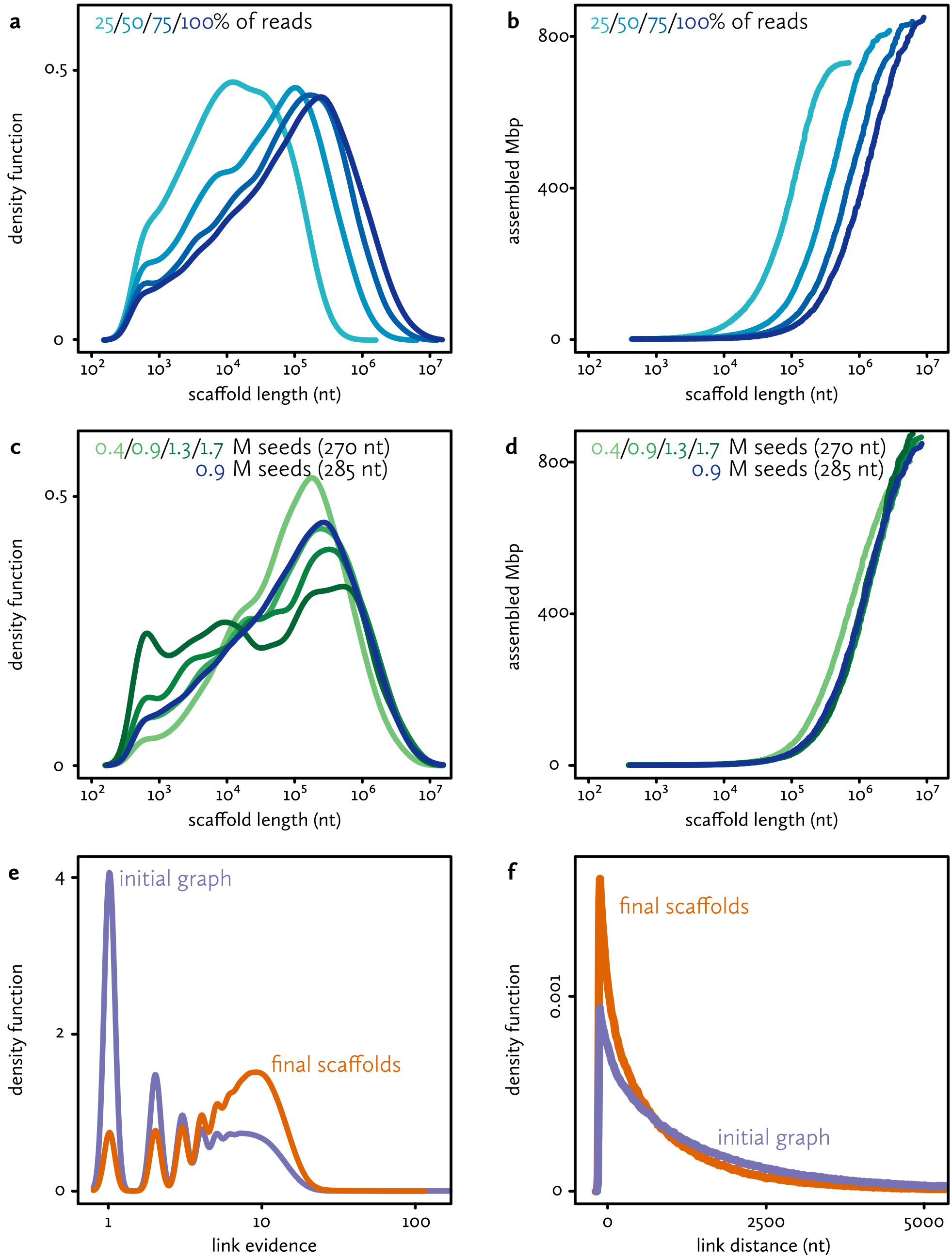

Figure 4 **Characteristics of the final assembly.**

**(a)** Size distribution of final scaffolds, based on 285 bp seeds.

Colours indicate alternative assembly runs, using subsets of the long read data.

**(b)** Cumulative size of the final scaffolds, sorted by size.

**(c)** and **(d)** Size distributions and cumulative size distributions for final scaffolds, based on both 270 and 285 bp seeds.

Colours indicate alternative assembly runs, using different seeds sets.

**(e)** Link evidence distribution in the initial graph (purple) and the final graph (orange) for the candidate assembly (285 bp seeds).

**(f)** Distances between seeds in the initial graph (purple) and the final graph (orange) for the candidate assembly (285 bp seeds).

## Assembly quality

In order to assess its completeness and structural correctness, we added nanopore sequence to the selected TULIP assembly and aligned it to the Illumina-based draft genome[2].
As a high-quality reference genome for the European eel is not yet available, such a comparison need take into account the possibility of error in either assembly. However, with appropriate caution, agreement between the assemblies – which are completely independent in both sequencing data and assembly algorithms – can confirm the integrity of both.

Figure 5a shows a full-genome alignment of the new (uncorrected) nanopore-based assembly to the 2012 draft2, based on best pairwise matches. This confirms that at this large scale, all sequence in the new assembly is also present in the older assembly. At first sight, the converse does not appear to be the case: the Illumina-based draft is 923 Mbp in size, and contains approximately 96 Mbp in scaffolds that have no reciprocal best match in the nanopore assembly (863.3 Mbp after sequence addition, see Supplementary Table S3). However, the non-matching sequences consist almost exclusively of very small scaffolds (mean/N50 664/987 bp). Since the Illumina-based draft assembly also contains 134 Mbp in gaps, these small scaffolds are plausibly sequences that could not be integrated correctly during the SSPACE scaffolding process[34, 35]. Both assemblies therefore roughly span the entire predicted genome of 860 Mbp.

Figure 5b–f show detailed alignments, based on the 5 largest nanopore scaffolds (6.1–8.9 Mbp uncorrected) and their best matches only. These alignments confirm that in this sample both assemblies are mostly collinear, with the smaller Illumina draft scaffolds usually aligning end-to-end on the larger TULIP scaffolds. Therefore, both presumably reflect the actual genomic organization. However, at this level of detail several structural incongruities between both assemblies also become apparent (indicated by arrowheads). For 16 scaffolds from the 2012 draft, only part of the sequence is present in the selected TULIP scaffolds. In other words, at these loci both assembly protocols made different choices, based on the available sequencing information.

We therefore examined the evidence for the decisions made by TULIP. For each discrepancy, we examined the local neighbourhoods in the initial nanopore-based seed graphs (as in Fig. 3). If a draft scaffold is correct, at the inconsistency there should be multiple alternatives for the TULIP algorithm to choose from (Supplementary Fig. S2). As these subgraphs (Supplementary Figs S3–S7) show, there is no evidence in the nanopore data for the older draft structure for any of the 16 cases examined. On the contrary, most local graph neighbourhoods appear relatively simple and support unambiguous scaffolding paths. The links at these suspect junctions are supported by at least two (average six) independent nanopore reads, which reduces the likelihood of accidental connections (caused by e.g. chimeric reads).

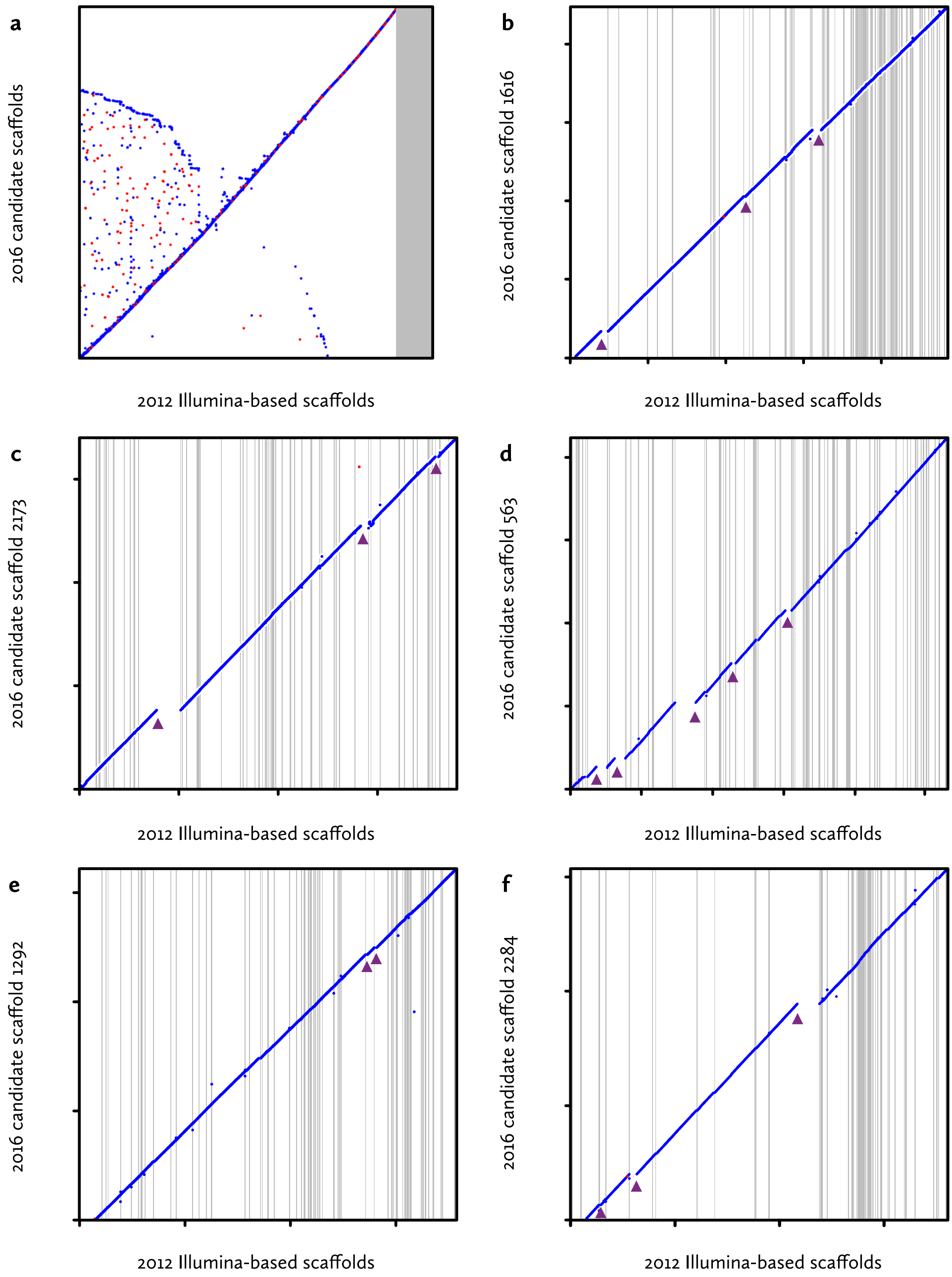

Figure 5 **Full-genome alignment of the final assembly.**

**(a)** The final uncorrected scaffolds (N50 = 1.19 Mbp, y-axis) were aligned to the 2012 *A. anguilla* assembly (N50 = 77.6 kbp, x-axis) using nucmer51 with minimum match length 100, filtered for best pairwise matches between scaffolds (delta-filter -1), and plotted using the mummerplot --layout option. The grey area corresponds to small scaffolds in the 2012 assembly that are not part of a best reciprocal match.

**(b–f)** More detailed alignments between the five largest nanopore scaffolds (y-axes) and their best matches in the 2012 draft assembly (x-axes). Grey vertical lines indicate scaffold boundaries. These figures were generated in R (version 3.3.1) based on mummerplot output. 2012 draft scaffolds with minimal contributions to the overall alignment were removed manually. Arrowheads indicate discrepancies between both assemblies.

Alternatively, the order of the draft scaffolds in the alignments already suggests which of the two assemblies is correct. If one of the 16 problematic scaffolds were to reflect the legitimate genome structure, this error in the new assembly would usually also affect the next aligning scaffold. However, in almost all cases, the neighbouring draft scaffold aligns end-to-end. This suggests that either the TULIP assembly intermittently features very large rearrangements that accidentally always end at draft scaffold boundaries, or that the draft scaffolds are occasionally misconstrued.

The distribution of draft scaffolds along the nanopore-based scaffolds reveals an interesting pattern. The distribution of draft scaffold length along the genome is clearly non-random, with some regions assembled into just a few large scaffolds, whereas other regions (often up to a Mbp in size) are highly fragmented into very small scaffolds. This indicates that using short-read technology, certain genomic features are intrinsically harder to assemble than using long reads.

Finally, we assessed the completeness of the nanopore assembly using BUSCO[36]. This method assumes complete assemblies to contain a high fraction of genes that are highly conserved in related species. From a set of 2586 common vertebrate genes, BUSCO was only able to recover 78 complete and 106 fragmented genes (3.0% and 4.1%, respectively). 92.9% of orthologues are missing from the nanopore assembly, indicating very poor completeness. In this case, however, this is a result of the sequence characteristics of ONT data.

## Sequence correction

Currently, the ONT platform does not yield reads of perfect sequence identity. Like with PacBio data, therefore, at some point in the assembly process the single-molecule-derived sequence needs to be corrected by extracting a consensus from multiple reads covering every genomic position. Here, we opted for a standalone post-assembly correction step with Racon, which extracts a consensus from nanopore reads[23]. As some positions in the assembly are based on a single nanopore read (Fig. 4e), in this case this correction may not be sufficient. Therefore, we subsequently corrected with Pilon, which extracts a consensus based on alignment of Illumina reads to the noisy sequence[37, 38].

To assess the changes made by these correction algorithms, we counted and compared the occurrence of 6-mers in the draft Illumina-based assembly, the uncorrected TULIP assembly, and after correction (Fig. 6). These frequencies reveal several expected patterns[17], specifically a slight underrepresentation of high CG content in Illumina-based sequence (draft and Pilon), and an underrepresentation of homopolymer sequence in nanopore-based sequence (TULIP and Racon). Overall, the correction steps bring the sequence similarity of the nanopore-based assembly closer to the Illumina-based draft, with the final corrected assembly having a high correlation to the draft (Fig. 6 lower left panel).

Sequence correction also has a strong positive impact on the BUSCO completeness assessment. As BUSCO relies on the prediction of gene structures, small artefactual deletions and insertions might cause it to miss genes. After correction with Racon, the BUSCO scores increased to 10.8% complete, 21.6% fragmented and 67.6% missing; correction with Pilon resulted in a further increase to 77.5% complete, 14.1% fragmented and 8.4% missing. An additional round of Pilon polishing resulted in a BUSCO assessment of 79.8% complete, 12.9% fragmented and 7.3% missing.

Sequence correction remains the most time-consuming stage of the assembly process, requiring 22 and 24 hours (on a single CPU) for Racon and Pilon, respectively (Supplementary Table S3). As TULIP bundles uncorrected scaffolds with its constituent nanopore reads, this process could still be sped up by parallelization, with individual scaffolds distributed over concurrent correction threads.

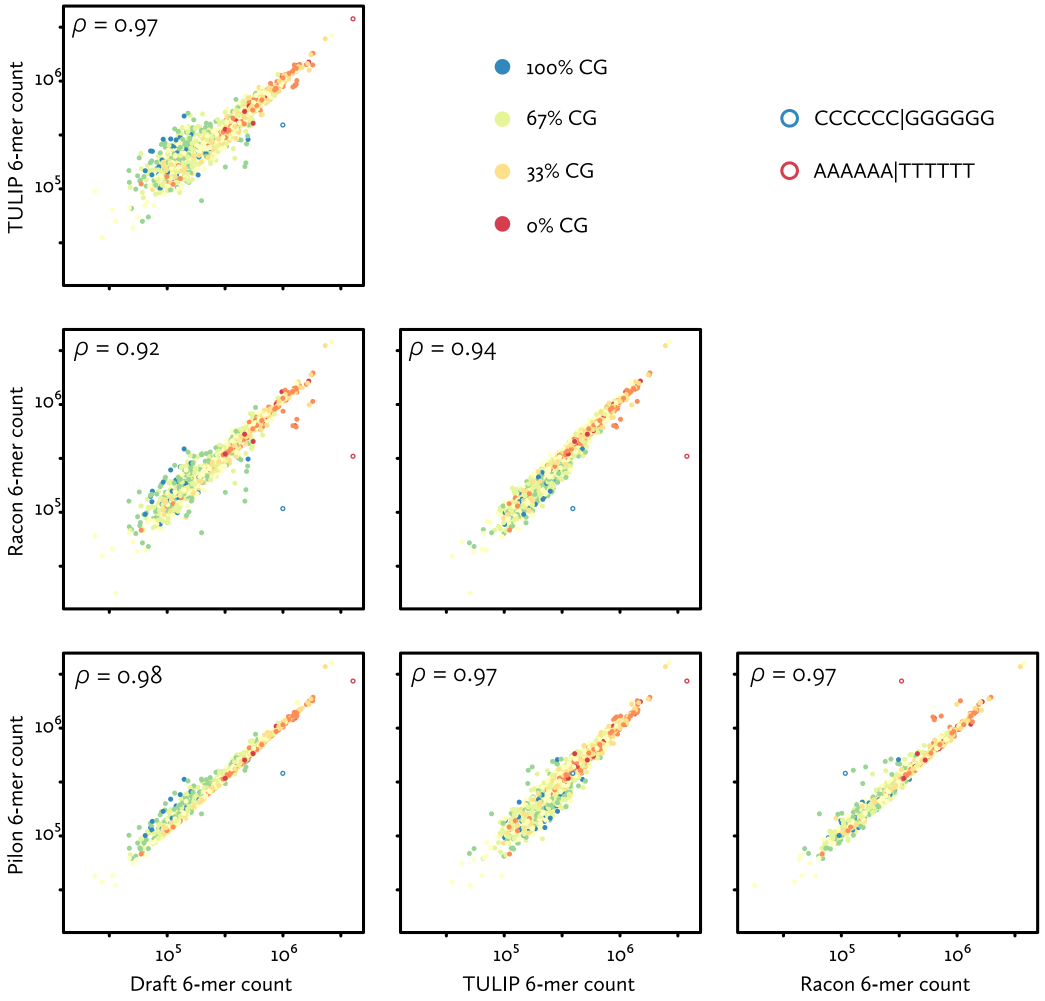

Figure 6 **Sequence identity in nanopore-based assemblies.**
The sequence similarity to the older draft of different stages of the nanopore assembly process (uncorrected TULIP, corrected by Racon[23], and additionally corrected by Pilon[37,38]) is illustrated by 6-mer frequency counts (generated using Jellyfish[46]). With every point a discrete 6-mer, colours indicate CG-content, and open circles indicate the two homo-6-mers. Scales are logarithmic. Also shown are Pearson correlation coefficients between the frequency distributions.

## — Discussion

In this study, we have evaluated whether it is possible to sequence a vertebrate genome using Oxford Nanopore long-read technology, and quickly assemble it by means of a relatively simple and lightweight procedure. Using our original TULIP methodology, we were able to assemble the 860 Mbp genome of the European eel using 18-fold nanopore coverage and sparse pre-selected Illumina reads in three and a half hours on a modest desktop computer. Including subsequent sequence correction, the entire process takes two days. This yields an assembly that is essentially complete and of high structural quality (Fig. 5).

One of the most striking outcomes of this eel genome sequencing effort is the close match between the genome size predicted from k-mer analysis (~860 Mbp) and the TULIP assembly (891.7 Mbp after corrections), and their distance from short-read-based assemblies. This can be explained either by the absence of a substantial fraction of the genome from the nanopore data or assembly, or by an artificially inflated genome size for the short-read assemblies. Full-genome alignment between both assemblies (Fig. 5a) suggests the latter phenomenon is at least partially responsible, as only tiny short-read scaffolds are absent from the long-read assembly. Furthermore, BUSCO analyses indicate the new assembly is approximately complete.

An analysis of the short-read *A. anguilla*[2] and *A. japonica*[35] assembly procedures implies that the scaffolding process, based on mate pair data, is responsible for the introduction of numerous gaps (Table 1). In addition, at the time we discarded a considerable fraction of the initial contigs, which was composed primarily of very small contigs that appeared to be artefactual (based on low read coverage or very high similarity to other contigs). Plausibly, such contigs – and the high residual fragmentation of these assemblies – are the result of the high levels of heterozygosity in these genomes (Supplementary Fig. S1).

Similar processes could also explain the even larger discrepancy between the predicted and assembled size of the recently published genome[39] of the American eel *A. rostrata* (Table 1). As European and American eels interbreed in the wild[40], a large difference in genome size is unlikely – although it could also provide an explanation for the observed limited levels of gene flow between the species[15].

The whole-genome alignments between the Illumina draft and the new nanopore-based assembly (Fig. 5) also serve to confirm the structural accuracy of both. In a representative sample (corresponding to of 4.2% of the genome), we observed 16 apparent assembly errors (Fig. 5b–f). In the absence of a high-quality reference, it is not straightforward to establish which assembly is correct. Our analyses, however, strongly suggest that in these cases the nanopore-based assembly is accurate. This is not unexpected: TULIP has access to far richer and more precise sequencing information than SSPACE, which had to rely on 2 × 36 bp mate pair data. Under such circumstances, a low number of incorrect joins between contigs is inevitable[41].
In fact, considering the fact that the SSPACE scaffolds analyzed in Fig. 5b–f consist of on the order of ten thousand very small contigs, a result with only 16 errors signifies better scaffolding performance than expected[41].

In other aspects, the TULIP assembly is likely to be suboptimal. By design, scaffolds that could be merged based on long reads remain separate if these reads do not share a fortuitous seed alignment in the correct position. Similarly, large repetitive regions in the genome, as well as (sub) telomeric repeats will not always contain frequent 285 bp islands of unique sequence, and hence could be absent from the assembly. Although counterintuitive, this should not pose a major problem for some extremely large genomes. Survey sequencing indicates that the 32 Gbp axolotl genome contains mostly unique sequence[29], as do many tulip genomes (C. Henkel, unpublished data).

The selection of sparse seeds by the user adds an unusual level of flexibility to the assembly process. In an early phase of this study, we opted for essentially randomly placed Illumina-based seed sequences. This choice was motivated by their very high sequencing identity, which aids alignment quality when working with noisy long reads. This strategy should work equally well with PacBio data or early, error-prone nanopore chemistries (i.e. R7.3).

The genome assembly generated here is a hybrid, incorporating two different sequencing technologies, three generations of nanopore sequencing, and two different animals. At the time, it was unavoidable to use a combination of multiple nanopore sequencing chemistries, as these rapidly replaced each other. Although the later R9 and R9.4 chemistries have better sequencing error profiles, they still retain structural biases that cannot be resolved by taking a consensus of nanopore data only (e.g. using Racon). In the final Pilon polishing stage, the nanopore data are therefore corrected using Illumina data obtained from a different eel specimen than used for nanopore sequencing. As the European eel is highly heterozygous (Table 2), in theory this generates a consensus between up to four different haplotypes. In practice, we expect this to have little influence on the quality of the final assembly, as the variation resulting from heterozygosity is much lower than the raw nanopore error rate. In other words, Pilon will treat SNPs and small indels not occurring in the Illumina data as sequencing errors to be corrected.

With the speed at which the quality of reads produced by the ONT platform is improving[18], it should soon be possible to avoid a hybrid assembly incorporating short reads altogether. A natural choice for seed sequences would then be the ends of long reads. Alternatively, seeds could be chosen to facilitate further sequence integration. If a high density genetic map is available for a species, map markers could serve as pre-ordered seeds. For example, with minor modifications, TULIP might be used to selectively add long read sequencing data only to single map marker bins (containing thousands of actual, unordered markers) resulting from a population sequencing strategy[42].

The bottleneck for such strategies lies in the interplay between marker density and nanopore read length, where the latter currently appears to be limited chiefly by DNA isolation protocols[43, 44]. Conceivably, in the near future, the problem of genome assembly from sequencing reads will all but disappear: abundant megabase-sized reads of high sequence identity are becoming possible, which should span the vast majority of recalcitrant regions in medium-sized genomes that remain a challenge to short- and medium-read technologies.

The fulfillment of such prophesies may still lie several years in the future. Therefore, we plan to further integrate and validate the candidate assembly generated here with long-range information obtained from optical mapping[45], in order to develop a high-quality reference genome for the troubled European eel.

## — Methods

### Eel samples

Two different European eels were used to generate the genome assembly. For all Illumina sequencing, a female specimen caught in Lake Veere, The Netherlands, was used. These data were previously used for the Illumina-based draft assembly[2]. For nanopore sequencing, a farmed female eel was obtained from Passie voor Vis, Sevenum, The Netherlands. As the European eel is a panmictic species[16], these sequenced eels belong to the same population. The experiments were approved by the animal ethical commission of Leiden University (DEC #13060), and carried out in accordance with the relevant guidelines and regulations.

### Genome size estimation and k-mer analyses

We used Jellyfish[46] version 2.2.6 to count k-mers in sequencing reads and assemblies. In order to estimate genome size, we obtained frequency histograms for 19- to 25-mers in raw Illumina sequencing data. Reads were truncated to a uniform length of 76 nt, except for *A. japonica*, for which we used 100 nt (the model did not converge for short lengths). For the American eel, which has been sequenced to much higher coverage than the European and Japanese species, we used a subset of the available data (NCBI Sequence Read Archive SRR2046741 and SRR2046672). Histograms were analyzed using the GenomeScope[33] website in order to obtain estimates for genome sizes, heterozygosity and duplication levels.

### Illumina seed selection

We selected unique seed sequences from 11.9 Gbp in sequence previously generated at 2 × 151 nt on an Illumina Hiseq 2000 (NCBI Sequence Read Archive SRR5235521). Pairs were merged using FLASh[47], requiring a minimum of 15 nt terminal overlaps, resulting in 29.16% merged fragments. In these, 25-mers were counted using Jellyfish. We used a custom script to filter out all fragments that contained 25-mers occurring over 25 times in the remaining data. This corresponds to a maximum occurrence of approximately 6.25× in the 860 Mbp genome. Finally, fragments were selected based on size (either 270 nt or 285 nt).

### MinION library preparation and sequencing

High MW chromosomal DNA was isolated from European eel blood and liver samples using a genomic tip 100 column according to the manufacturer's instructions (Qiagen). For each nanopore sequencing library, we used 2–3 µg genomic DNA, approximately twice the recommended quantity. In this way, we compensated for the decreased molar quantities of DNA ends at increased fragment lengths (see below).

First the DNA was sequenced on R7.3 flow cells. Subsequently multiple R9 and R9.4 flow cells were used to sequence the DNA. For R7.3 sequencing runs we prepared the library using the SQK-MAP006 kit from Oxford Nanopore Technologies. Briefly, high molecular weight DNA was sheared with a g-TUBE (Covaris) to an average fragment length of 20 kbp. The sheared DNA was repaired using the FFPE repair mix according to the manufacturer's instructions (New England Biolabs, Ipswich, USA).

After cleaning up the DNA with an extraction using a ratio of 0.4:1 Ampure XP beads to DNA the DNA ends were polished and an A overhang was added with the the NEBNext End Prep Module and again cleaned up with an extraction using a ratio of 1:1 Ampure XP beads to DNA the DNA prior to ligation. The adaptor and hairpin adapter were ligated using Blunt/TA Ligase Master Mix (New England Biolabs). The final library was prepared by cleaning up the ligation mix using MyOne C1 beads (Invitrogen).

To prepare 2D libraries for R9 sequencing runs we used the SQK-NSK007 kit from Oxford Nanopore Technologies. The procedure to prepare a library with this kit is largely the same as with the SQK-MAP006 kit. 1D library preparation was done with the SQK-RAD001 kit from Oxford Nanopore Technologies. In short, high molecular weight DNA was tagmented with a transposase. The final library was prepared by ligation of the sequencing adapters to the tagmented fragments using the Blunt/TA Ligase Master Mix (New England Biolabs). Library preparation for R9.4 sequencing runs was done with the SQK-LSK108 and the SQK-RAD002 kits from Oxford Nanopore Technologies. The procedure to prepare libraries using the SQK-RAD002 kit was the same as for the SQK-RAD001 kit. For SQK-LSK108 the procedure was essentially the same as for SQK-NSK007 except that only adapters and no hairpins were ligated to the DNA fragments. As a consequence the final purification step was done using Ampure XP beads instead of MyOne C1 beads. Libraries for R7.3 and R9 flow cells were directly loaded on the flow cells. To load the library on the R9.4 flow cell the DNA fragments were first bound to beads which were then loaded on the flow cell.

The MinKNOW software was used to control the sequencing process and the read files were uploaded to the cloud based Metrichor EPI2ME platform for base calling. Base called reads were downloaded for further processing and assembly.

## Nanopore read alignment

From the base called read files produced by the Metrichor EPI2ME platform sequence files in FASTA format were extracted using the R-package poRe version 0.17 (ref. 48). We used BWA-MEM[49] (version 0.7.15-r1140) to align nanopore reads to selected seeds, using specific settings for each nanopore chemistry. The built-in -x ont2d setting (-k 14 -W 20 -r 10 -A 1 -B 1 -O 1 -E 1 -L 0) is too tolerant for newer chemistries. We therefore optimized alignment settings (-k and -W only) on small subsets to yield the highest recall (number of aligning reads) at the highest precision (number of seeds detected/number of alignments). With all other settings as before, this yielded the following parameters: -k 14 -W 45 (R7.3 2D); -k 16 -W 50 (R9 1D); -k 19 -W 60 (R9 2D); -k 16 -W 60 (R9.4 1D).

## Genome assembly using TULIP

Currently, TULIP consists of two prototype scripts in Perl: tulipseed.perl and tulipbulb.perl (version 0.4 ‘European eel’). The tulipseed script constructs the seed graph based on input SAM files and a set seed length, and outputs a simplified graph and seed arrangements (scaffold models). tulipbulb adds seed and long read sequence to the scaffolds, and exports either a complete set of uncorrected scaffolds, or for each scaffold two separate files: the uncorrected sequence, and a FASTA ‘bundle’ consisting of all long reads associated with that scaffold.

For each scaffold, we used the long read bundle and Illumina data to polish it according to ONT guidelines (https://github.com/nanoporetech/ont-assembly-polish). We first corrected nanopore-derived scaffolds with nanopore data using Racon[22], based on alignments produced by Graphmap[50] version 0.3.0. Ultimately Racon sequence correction is performed by SPOA51, which is a partial order alignment algorithm that generates consensus sequences.

Subsequently, we used previously generated[2] Illumina data (NCBI Sequence Read Archive SRR5235521– SRR5235523), trimmed to Phred 30 quality values (using Sickle version 1.33, https://github.com/najoshi/sickle) in a second correction step using Pilon (version 1.21), an integrated software tool for assembly improvement[37, 38]. Pilon uses evidence from the alignment between short-read data and Racon-corrected scaffolds to identify events that are different in the draft genome compared to the support of short-read data.

All genome assembly steps and analyses were performed on a desktop computer equipped with an Intel Xeon E3-1241 3.5 GHz processor, in a virtual machine (Oracle VirtualBox version 4.3.26) running Ubuntu 16.04 LTS with 28 GB RAM and 4 processor threads available. For the final candidate assembly, the TULIP scripts required a maximum of 4.4 GB RAM.

## Genome alignment

Uncorrected scaffolds were aligned against the 2012 scaffolds using nucmer[52] version 3.23, with settings --maxmatch and --minmatch 100, filtered for optimal correspondence (delta-filter -1), and visualized using mummerplot (with the --layout option). The five largest scaffolds were likewise aligned against the 2012 scaffolds, but with settings encouraging longer alignments ( --breaklen 1000 and --minmatch 25) and not filtered. The 285 nt seeds were aligned against the 2012 draft scaffolds using BWA-MEM with default settings.

## BUSCO assembly assessment

The completeness of the genome assemblies was tested with BUSCO[36] (version 3.0.0), which tries to find orthologues of a curated dataset of near-universal genes in new assemblies. A more complete assembly will result in a higher percentage of genes retrieved. As the European eel is a primitive teleost, we used the vertebrate-specific orthologue catalogue (vertebrata_odb9, creation date 13-2-2016, 2586 genes) instead of actinopterygii_odb9, which is based predominantly on the genome sequences of advanced teleosts.

## Data availability

The nanopore sequencing data are available in the European Nucleotide Archive (accession number PRJEB20018). The Racon- and Pilon-corrected candidate assembly is available at http://www.eelgenome.com.  The TULIP-scripts are available at https://github.com/Generade-nl

## Acknowledgements

We are grateful to the DUPAN Foundation for making this project possible, as well as to Rosemary Dokos and Oliver Hartwell at Oxford Nanopore for technical support and encouragement.
This project was funded by grants from the DUPAN Foundation for sustainable eel farming and fishing, the Dutch Ministry of Economic Affairs (to A.P.P., KB-21-001-001), the Austrian Science Foundation (to B.P., FWF P26363-B25), the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Actions: Innovative Training Network IMPRESS, grant agreement No 642893 (to F.-A.W.), ZonMW/DTL (to C.V.H.), and local funds from CNRS (to S.D.) and Generade, the Leiden Centre of Expertise in Genomics (to C.V. H.).

## Author Contributions

H.J.J., S.D., F.-A.W., W.S., A.K., A.P.P., B.P., H.P.S., G.E.V.D.T., R.P.D. and C.V.H. conceived the research. R.P.D. coordinated the project. H.J.J. and S.A.J.-R. performed sequencing, M.L. and C.V.H. assembled the genome, H.J.J., R.P.D. and C.V.H. analyzed the data. H.J.J., M.L., R.P.D. and C.V.H. wrote the paper with input from all other authors.

## Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-07650-6

## Competing Interests

H.J.J. and C.V.H. are members of the Nanopore Community, and have previously received flow cells free of charge (used for some of the R7.3 data of this project), as well as travel expense reimbursements from Oxford Nanopore Technologies.

## References

1. Coppe, A. et al. Sequencing, *de novo* annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. BMC Genomics 11, 635 (2010).

2. Henkel, C. V. et al. Primitive duplicate Hox clusters in the European eel's genome. PLoS One 7, e32231 (2012).

3. Pujolar, J. M. et al. Surviving in a toxic world: transcriptomics and gene expression profiling in response to environmental pollution in the critically endangered European eel. BMC Genomics 13, 507 (2012).

4. Minegishi, Y., Henkel, C. V., Dirks, R. P. & van den Thillart, G. E. Genomics in eels – towards aquaculture and biology. Mar Biotechnol (NY) 14, 583–590 (2012).

5. IUCN Red List. doi:10.2305/IUCN.UK.2014-1.RLTS.T60344A45833138.en (2014).

6. Ager-Wick, E. et al. The pituitary gland of the European eel reveals massive expression of genes involved in the melanocortin system. PLoS One 8, e77396 (2013).

7. Dirks, R. P. et al. Identification of molecular markers in pectoral fin to predict artificial maturation of female European eels (*Anguilla anguilla*). Gen Comp Endocrinol 204, 267–276 (2014).

8. Churcher, A. M. et al. Deep sequencing of the olfactory epithelium reveals specific chemosensory receptors are expressed at sexual maturity in the European eel *Anguilla anguilla*. Mol Ecol 24, 822–834 (2015).

9. Burgerhout, E. et al. Changes in ovarian gene expression profiles and plasma hormone levels in maturing European eel (*Anguilla anguilla*); biomarkers for broodstock selection. Gen Comp Endocrinol 225, 185–196 (2016).

10. Pelster, B., Schneebauer, G. & Dirks, R. P. Anguillicola crassus infection significantly affects the silvering related modifications in steady state mRNA levels in gas gland tissue of the European eel. Front Physiol 7, 175 (2016).

11. Pujolar, J. M. et al. A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. Mol Ecol Resour 13, 706–714 (2013).

12. Pasquier, J. et al. Multiple kisspeptin receptors in early osteichthyans provide new insights into the evolution of this receptor family. PLoS One 7, e48931 (2012).

13. Maugars, G. & Dufour, S. Demonstration of the coexistence of duplicated LH receptors in teleosts, and their origin in ancestral actinopterygians. PLoS One 10, e0135184 (2015).

14. Morini, M. et al. Duplicated leptin receptors in two species of eel bring new insights into the evolution of the leptin system in vertebrates. PLoS One 10, e0126008 (2015).

15. Jacobsen, M. W. et al. Genomic footprints of speciation in Atlantic eels (*Anguilla anguilla* and *A. rostrata*). Mol Ecol 23, 4785–4798 (2014).

16. Pujolar, J. M. et al. Genome-wide single-generation signatures of local selection in the panmictic European eel. Mol Ecol 23, 2514–2528 (2014).

17. Ip, C. L. et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res 4, 1075 (2015).

18. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17, 239 (2016).

19. Tyson, J. R. et al. Whole genome sequencing and assembly of a Caenorhabditis elegans genome with complex genomics rearrangements using the MinION sequencing device. BioRxiv, doi:10.1101/099143 (2017).

20. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 1; referees: 2 approved with reservations]. F1000Res 6, 100 (2017).

21. Koren, S. et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14, R101 (2013).

22. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat Methods 12, 733–735 (2015).

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

24. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13, 1050–1054 (2016).

25. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27, 722–736 (2017).

26. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & Tse, D. N. HINGE: long-read assembly achieves optimal repeat resolution. Genome Res 27, 747–756 (2017).

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

28. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep 6, 31900 (2016).

29. Keinath, M. C. et al. Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. Sci Rep 5, 16413 (2015).

30. Biscotti, M. A. et al. The lungfish transcriptome: a glimpse into molecular evolution events at the transition from water to land. Sci Rep 6, 21571 (2016).

31. Zonneveld, B. J. The systematic value of nuclear genome size for all species of Tulipa L. (Liliacaeae). Plant Syst Evol 281, 217–245 (2009).

32. Li, X. & Waterman, M. S. Estimating the repeat structure and length of DNA sequences using l-tuples. Genome Res 13, 1916–1922(2003).

33. Vuture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, doi:10.1093/bioinformatics/btx153 (2017).

34. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579 (2011).

35. Henkel, C. V. et al. First draft genome sequence of the Japanese eel. *Anguilla japonica*. Gene 511, 195–201 (2012).

36. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).

37. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963 (2014).

38. Goodwin, S. et al. Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. Genome Res 25, 1–7 (2015).

39. Pavey, S. A. et al. Draft genome of the American eel (*Anguilla rostrata*).
Mol Ecol Resour, doi:10.1111/1755-0998.12608 (2016).

40. Albert, V., Jónsson, B. & Bernatchez, L. Natural hybrids in Atlantic eels (*Anguilla anguilla*, *A. rostrata*): evidence for successful reproduction and fluctuating abundance in space and time. Mol Ecol 15, 1903–1916 (2006).

41. Hunt, M., Newbold, C., Berriman, M. & Otto, T. D. A comprehensive evaluation of assembly scaffolding tools.
Genome Biol 15, R42(2014).

42. Chapman, J. A. et al. A whole-genome shotgun approach for assembling and anchoring the hexaploidy bread wheat genome. Genome Biol 16, 26 (2015).

43. Urban, J. M., Bliss, J., Lawrence, C. E. & Gerbi, S. A. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. BioRxiv, doi:10.1101/019281 (2015).

44. Datema, E. et al. The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. BioRxiv, doi:10.1101/084772 (2016).

45. Mostovoy, Y. et al. A hybrid approach for *de novo* human genome sequence assembly and phasing.
Nat Methods 13, 587–590 (2016).

46. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.
Bioinformatics 27, 764–770 (2011).

47. Magoc, T. & Salzberg, S. FLASH: Fast length adjustment of short reads to improve genome assemblies.
Bioinformatics 27, 2957–2963 (2011).

48. Watson, M. et al. poRe: an R package for the visualization and analysis of nanopore sequencing data.
Bioinformatics 31, 114–115 (2015).

49. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
Bioinformatics 26, 589–95 (2010).

50. Sović, I. et al. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun 7, 11307 (2016).

51. Lee, C. Generating consensus sequences from partial order multiple sequence alignment graphs. Bioinformatics 19, 999–1008 (2003).

52. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol 5, R12 (2004).

— Chapter 5

# Microbial diversity characterization of seawater in a pilot study using Oxford Nanopore Technologies long-read sequencing

An abbreviated version was published at BMC Res Notes (2021) 14:42

M. Liem[1]*, T. Regensburg-Tuïnk[1], C. Henkel[2], H. Jansen[3] and H. Spaink[1]

[1] Institute Biology, Leiden University, the Netherlands

[2] Norwegian University of Life Sciences (NMBU)

[3] Future Genomics Technologies, the Netherlands

*corresponding author:
Michael Liem
Sylviusweg 72
2333 BE Leiden
the Netherlands
+31 71 527 5000
m.liem@biology.leidenuniv.nl

## — Abstract

### Objective

Currently the majority of non-culturable microbes in sea water are yet to be discovered, Nanopore offers a solution to overcome the challenging tasks to identify the genomes and complex composition of oceanic microbiomes. In this study we evaluate the utility of Oxford Nanopore Technologies (ONT) sequencing to characterize microbial diversity in seawater from multiple locations. We compared the microbial species diversity of retrieved environmental samples from two different locations and time points.

### Results

With only three ONT flow cells we were able to identify thousands of organisms, including bacteriophages, from which a large part at species level. It was possible to assemble genomes from environmental samples with Flye. In several cases this resulted in >1 Mbp contigs and in the particular case of a *Thioglobus singularis* species it even produced a near complete genome. k-mer analysis reveals that a large part of the data represents species of which close relatives have not yet been deposited to the database. These results show that our approach is suitable for scalable genomic investigations such as monitoring oceanic biodiversity and provides a new platform for education in biodiversity.

Keywords Metagenomics, Oxford Nanopore Technologies, MinION sequencing, oceanic microbiome, k-mer analysis, genome assembly

## — Introduction

Although marine microbes have been studied for multiple decades there is still little knowledge on species diversity in the largest ecological environments of our planet[1-3]. Current database collections are estimated to represent <5% of oceanic microbial communities[4]. Seawater contains many non-culturable organisms, hence to understand its microbial ecology we need to collect sequencing data from DNA samples obtained directly from the environment.

Large-scale metagenomics analyses of seawater have been performed already since 2004 showing remarkable species diversity[5]. However, even with availability of abundant sequencing technology resources a complete understanding on the entire diversity remains a challenging task. This is due to, among others, vast water volumes and huge amounts of microbe communities, which through temporal and spatial dynamics contribute to the existence of a near infinite number of ecosystems. Recent studies focussing on marine biodiversity show that a variety of sediments harbour different ecosystems that are particularly extreme in deep ocean environments. There have been many exploratory studies of harnessing marine microorganism for the production of bioactive compounds, with versatile medicinal, industrial, or agricultural applications[6].

Microbial diversity characterization has primarily relied on traditional high-throughput short-read sequencing methods, such as Illumina[7-12] or 454 sequencing[5]. Even though Pacific Biosciences single-molecule long-read sequencing has been used to catalogue the diversity of coral-associated microbial communities, these studies relied on amplification and 16S rRNA homology to position microbes taxonomically[5, 7, 13, 9-11, 14]. Amplification, however, introduces biases that results in over- and under-representation of particular species. Additionally, in some cases 16S rRNA identification fails to characterize microbial diversity due to variability in the 16S region[15], – for example, previous studies revealed that some universal primers have strong biases against the detection of pelagic bacteria (SAR11 group) and archaea[4].

Hence 16S-based methods appear ineffective at comprehensively characterizing complex metagenomics samples such as from seawater. Furthermore, traditional 16S rRNA identification is limited to the detection of microbe presence and does not yield further functional insights about the organism. And finally, high-throughput short read sequencing methods require large scale infrastructure including sequencers and laboratories.

In this pilot study we evaluate the utility of Oxford Nanopore Technologies (ONT) sequencing to characterize microbial diversity in seawater.
ONT sequencing generates on average 10 Kbp reads, theoretically without upper limit, and bypasses the necessity of amplification. Our strategy aims to classify microbial diversification directly from environmental samples (two different oceanic locations were chosen) with minimal computational and financial cost over a relatively short time span. This will facilitate future scalable investigations such as monitoring oceanic biodiversity and the time and space dynamics these microbes are subject to.

## — Results

### Sample collection, data quality control and verification of microbial content

We collected samples from coastal regions of both the Atlantic Ocean (west part of the English Channel – Roscoff, France, August 2017) and the south part of the North Sea (Wassenaarseslag, the Netherlands, July 2017 and August 2018). From here on, we refer to these as samples 1, 2 and 3. MinION 48-hour sequencing runs on every sample resulted in three datasets with mean read lengths that range between 1,511 and 7,983 bp (Table 2). Our read length distributions indicate relatively suboptimal DNA samples that resulted in shorter reads (Figure 1) compared to ONT read length averages of laboratory cultures. This is particularly apparent for sample 1. The error rate expressed in PHRED indicates similar quality for the three runs, our average qualities fluctuate around PHRED 12 that stands for <10% error per read on average.

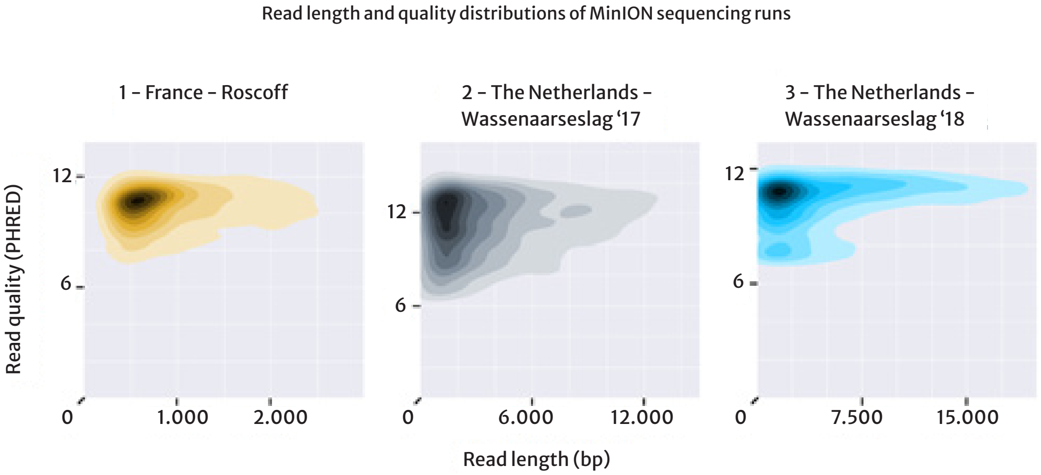

Figure 1 Read length and quality distributions of 48-hour run sequencing data for sample 1, 2 and 3 (from left to right). Mean read lengths vary from 1,511 up to 7,983 bp with similar base call qualities (around PHRED 12). Plots are based on NanoPlot plotting[23]

To assess quality of the data we analysed homologues sequences of the three longest reads for all three data sets. The results (Table 1) show that several of these reads are representative of bacterial species that were found to be dominant by the OneCodex analyses. One of the reads (France – Read ID 1) also showed that we have identified a representative of a bacteriophage of *Pelagibacter*. The limited coverage of the homologue genes indicates that we have identified a rather distant new relative of the published bacteriophage.

Table 1 **Blast alignment of longest raw sequencing reads.**
Sample) time and location of seawater samples, Read ID) read length identifier sorted from longest to smallest, Query length) the length of the read, Best hits*)* criteria for best hit; largest query coverage with highest identity and published study, Cov) alignment percentage that reads cover the reference, ID) alignment identity between query and reference, Ref length) length of the reference sequence.

| Sample | Read ID | Query length (Kbp) | Best hits * | Cov (%) | ID (%) | Ref length (Kbp) |
|---|---|---|---|---|---|---|
| Fr. | 1 | 50 | *Pelagibacter* phage HTVC008M[29] | 2 | 78 | 147 |
| Fr. | 2 | 46 | *Candidatus Pelagibacter* sp. FZCC0015<br>[CP031125] | 3 | 68 | 1,364 |
| Fr. | 3 | 45 | *Halioglobus pacificus* strain RR3-57<br>[CP019046] | 9 | 69 | 4,847 |
| NL '17 | 1 | 155 | *Brassica oleracea* HDEM<br>[LR031920] | 3 | 68 | 113 |
| NL '17 | 2 | 107 | No hit – repetitive stretch | | | |
| NL '17 | 3 | 78 | *Halioglobus japonicus* strain NBRC 107739<br>[CP019450] | 2 | 69 | 4,085 |
| NL '18 | 1 | 161 | *Flavobacterium columnare*[31] | 28 | 68 | 3,329 |
| NL '18 | 2 | 149 | *Clostridium tetani* strain Harvard 49205<br>[CP035787] | <1 | 69 | 2,807 |
| NL '18 | 3 | 139 | *Micromonas* sp. RCC1109 virus MpV1[30] | 23 | 74 | 184 |

To confirm that our double filtering method indeed selects for microbial DNA we have used 16S rRNA primers that are known to identify a wide range of microbial genomes. FastPCR aligns the currently 'best available' 16S rRNA primer sequences[25] to raw sequencing data and shows microbial content in all three raw sequencing datasets. We found 23, 178 and 188 hits aligning both forward and reverse primers that span between 420 and 470 bp (Table 2). These hits have a minimum of 80% alignment identity and ranged up to 100% matches. Blast searches of regions that have <80% sequence identity did not result in hits originating from 16S rRNA hence do not contribute to the identification of microbial content and have been omitted.

Table 2 Raw sequencing data statistics of sample 1,2 and 3

| Statistics | France (1) | The Netherlands '17 (2) | The Netherlands '18 (3) |
|---|---|---|---|
| *Reads* | 370,371 | 1,316,823 | 225,200 |
| *Bases* | 559,696,414 | 6,350,530,291 | 1,797,851,809 |
| *Mean length (bp)* | 1,511 | 4,822 | 7,983 |
| *Max length (bp)* | 49,807 | 155,979 | 161,655 |
| *16S reads* | 23 | 178 | 188 |

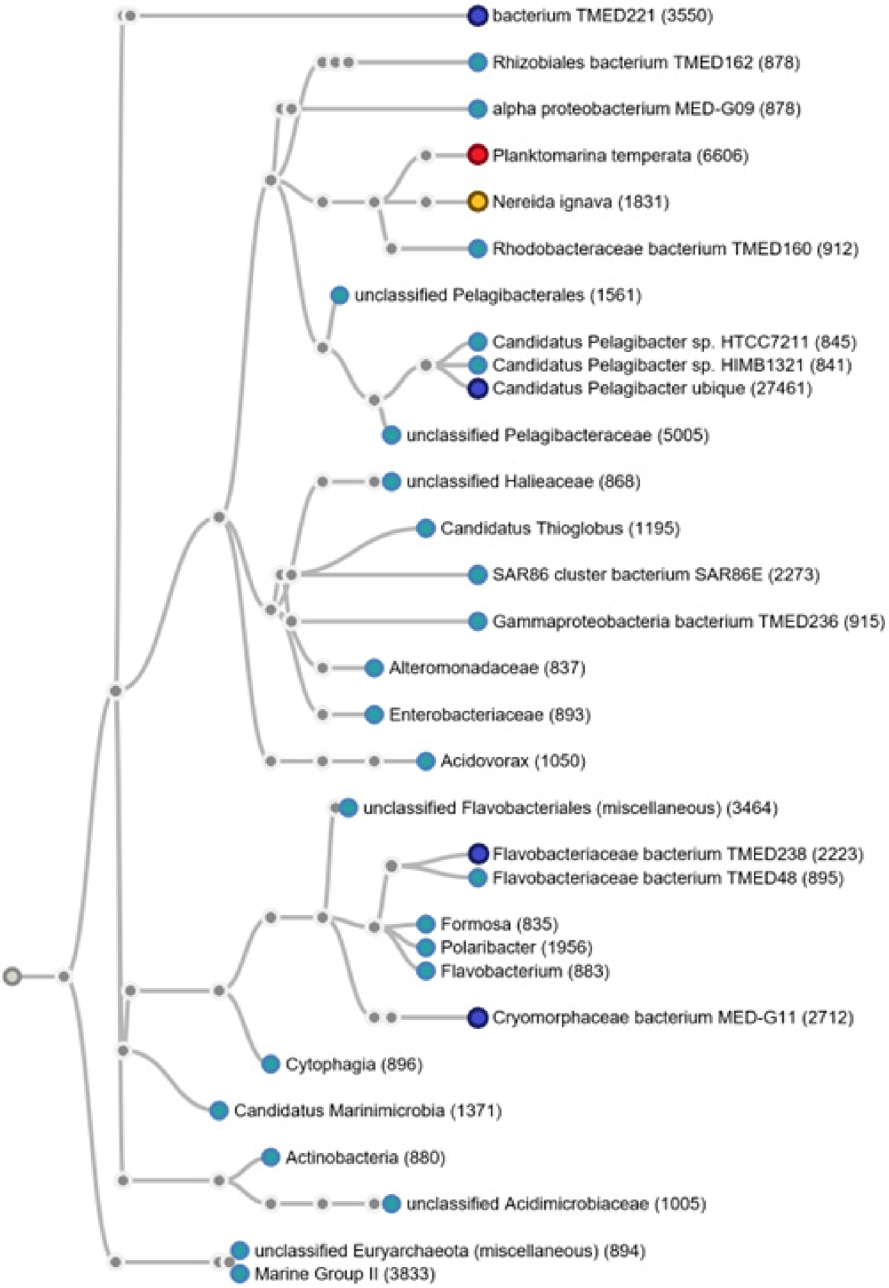
bacterium TMED221 (3550)
Rhizobiales bacterium TMED162 (878)
alpha proteobacterium MED-G09 (878)
Planktomarina temperata (6606)
Nereida ignava (1831)
Rhodobacteraceae bacterium TMED160 (912)
unclassified Pelagibacterales (1561)
Candidatus Pelagibacter sp. HTCC7211 (845)
Candidatus Pelagibacter sp. HIMB1321 (841)
Candidatus Pelagibacter ubique (27461)
unclassified Pelagibacteraceae (5005)
unclassified Halieaceae (868)
Candidatus Thioglobus (1195)
SAR86 cluster bacterium SAR86E (2273)
Gammaproteobacteria bacterium TMED236 (915)
Alteromonadaceae (837)
Enterobacteriaceae (893)
Acidovorax (1050)
unclassified Flavobacteriales (miscellaneous) (3464)
Flavobacteriaceae bacterium TMED238 (2223)
Flavobacteriaceae bacterium TMED48 (895)
Formosa (835)
Polaribacter (1956)
Flavobacterium (883)
Cryomorphaceae bacterium MED-G11 (2712)
Cytophagia (896)
Candidatus Marinimicrobia (1371)
Actinobacteria (880)
unclassified Acidimicrobiaceae (1005)
unclassified Euryarchaeota (miscellaneous) (894)
Marine Group II (3833)

Figure 2 **Taxonomic tree on a subset of the data generated from sample 1 data.**
Every node stands for a taxonomical ID that is supported with at least 831 reads. In red the most abundant species present in all three samples. Dark blue nodes together with the red node highlight the top-5 most abundantly present species in this sample. The yellow node indicates the most prominent species difference between the two locations.

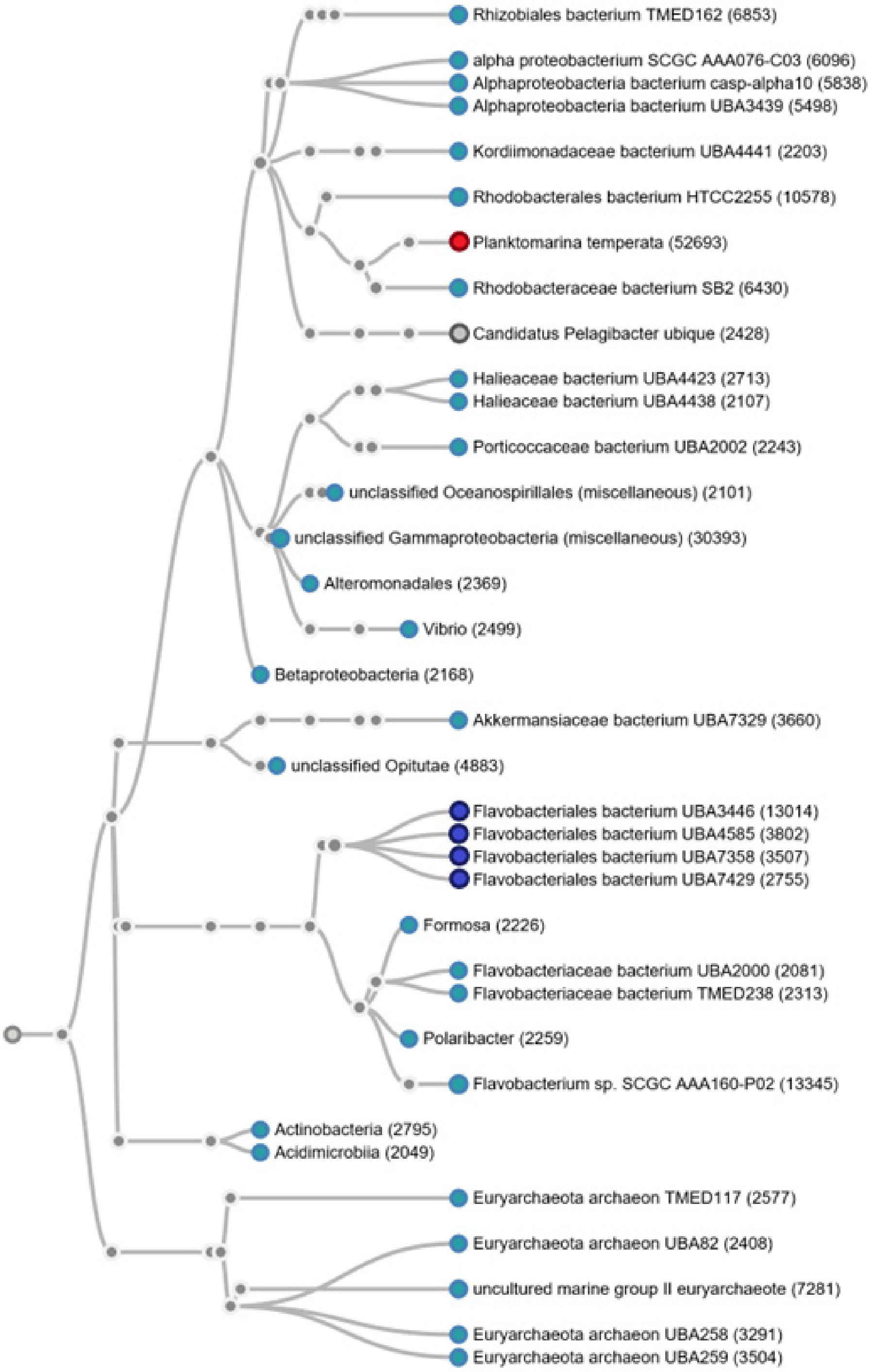
Rhizobiales bacterium TMED162 (6853)
alpha proteobacterium SCGC AAA076-C03 (6096)
Alphaproteobacteria bacterium casp-alpha10 (5838)
Alphaproteobacteria bacterium UBA3439 (5498)
Kordiimonadaceae bacterium UBA4441 (2203)
Rhodobacterales bacterium HTCC2255 (10578)
Planktomarina temperata (52693)
Rhodobacteraceae bacterium SB2 (6430)
Candidatus Pelagibacter ubique (2428)
Halieaceae bacterium UBA4423 (2713)
Halieaceae bacterium UBA4438 (2107)
Porticoccaceae bacterium UBA2002 (2243)
unclassified Oceanospirillales (miscellaneous) (2101)
unclassified Gammaproteobacteria (miscellaneous) (30393)
Alteromonadales (2369)
Vibrio (2499)
Betaproteobacteria (2168)
Akkermansiaceae bacterium UBA7329 (3660)
unclassified Opitutae (4883)
Flavobacteriales bacterium UBA3446 (13014)
Flavobacteriales bacterium UBA4585 (3802)
Flavobacteriales bacterium UBA7358 (3507)
Flavobacteriales bacterium UBA7429 (2755)
Formosa (2226)
Flavobacteriaceae bacterium UBA2000 (2081)
Flavobacteriaceae bacterium TMED238 (2313)
Polaribacter (2259)
Flavobacterium sp. SCGC AAA160-P02 (13345)
Actinobacteria (2795)
Acidimicrobiia (2049)
Euryarchaeota archaeon TMED117 (2577)
Euryarchaeota archaeon UBA82 (2408)
uncultured marine group II euryarchaeote (7281)
Euryarchaeota archaeon UBA258 (3291)
Euryarchaeota archaeon UBA259 (3504)

Figure 3 **A subset of the data set from sample 2, every node is supported with minimally 2048 reads.** The red node indicates the most abundant species over all three datasets, together with dark blue nodes it comprises the top-5 most abundant species in this dataset. Particularly underrepresented is species Candidatus Pelagibacter (grey node) compared to sample 1 and 3.

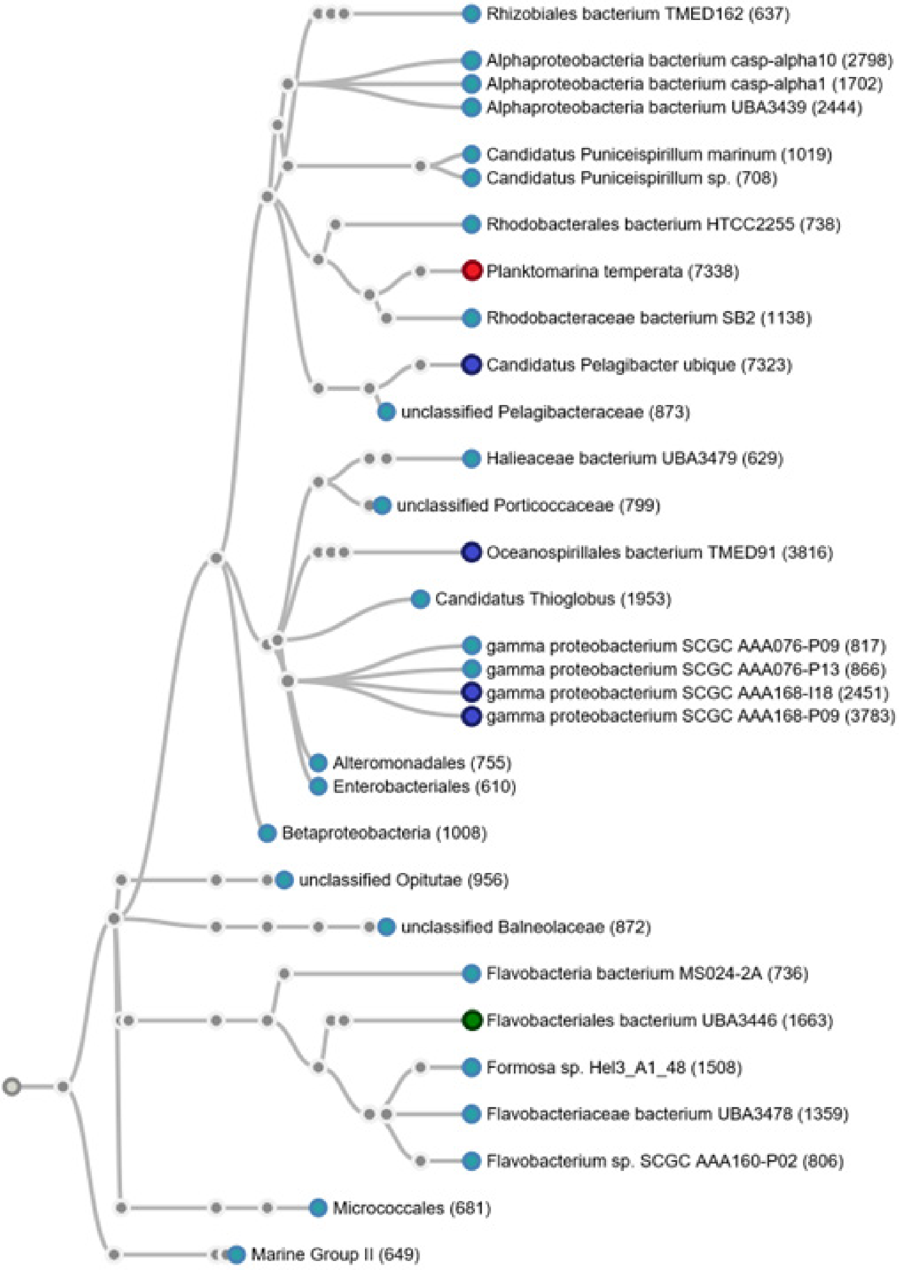
Rhizobiales bacterium TMED162 (637)
Alphaproteobacteria bacterium casp-alpha10 (2798)
Alphaproteobacteria bacterium casp-alpha1 (1702)
Alphaproteobacteria bacterium UBA3439 (2444)
Candidatus Puniceispirillum marinum (1019)
Candidatus Puniceispirillum sp. (708)
Rhodobacterales bacterium HTCC2255 (738)
Planktomarina temperata (7338)
Rhodobacteraceae bacterium SB2 (1138)
Candidatus Pelagibacter ubique (7323)
unclassified Pelagibacteraceae (873)
Halieaceae bacterium UBA3479 (629)
unclassified Porticoccaceae (799)
Oceanospirillales bacterium TMED91 (3816)
Candidatus Thioglobus (1953)
gamma proteobacterium SCGC AAA076-P09 (817)
gamma proteobacterium SCGC AAA076-P13 (866)
gamma proteobacterium SCGC AAA168-I18 (2451)
gamma proteobacterium SCGC AAA168-P09 (3783)
Alteromonadales (755)
Enterobacteriales (610)
Betaproteobacteria (1008)
unclassified Opitutae (956)
unclassified Balneolaceae (872)
Flavobacteria bacterium MS024-2A (736)
Flavobacteriales bacterium UBA3446 (1663)
Formosa sp. Hel3_A1_48 (1508)
Flavobacteriaceae bacterium UBA3478 (1359)
Flavobacterium sp. SCGC AAA160-P02 (806)
Micrococcales (681)
Marine Group II (649)

Figure 4 **Taxonomic tree on a subset of sequencing data from sample 3, every node is supported with at least 588 reads.** Again the red node indicates the overall most abundant species, and together with dark blues nodes they form the top-5 most abundant species for this dataset. Compared to the year before Flavobacteriales bacterium is underrepresented (green node).

Using OneCodex[26] we generated classification trees for the three datasets. These are built from raw sequencing data and indicate the taxonomic relation between the detected microbial classes. This relation is based on taxonomic identifiers (taxids) provided by the NCBI taxonomy database. For visualization purposes these taxonomic trees are subsets of the complete classifications: every node is supported with a minimum threshold of 831, 2,048 and 588 reads for samples 1, 2 and 3, respectively.

Despite the fact that a large part of all three datasets could not be classified (47%, 69% and 38% for sample 1, 2 and 3, respectively), all taxonomic trees highlight the complexity of microbial communities present at a single site. None of our three datasets reveal an overall dominant species, with the largest differences between samples microbes that appear at low abundances. However 4.46% (sample 1), 15.66% (sample 2) and 7.82% (sample 3) of classified reads belong to *Planktomarina temperata*, which is therefore the most abundant species present in the three data sets combined (Figure 2, Figure 3 and Figure 4, red nodes).

The top-5 most abundant species in sample 1 are: *Candidatus Pelagibacter ubique* (9.31% of Proteobacteria), *bacterium* TMED221 (8.61% of unclassified bacteria), *Flavobacteriaceae bacterium* TMED238 (4.48% of the FCB group), *Planktomarina temperata* (4.46% of Proteobacteria) and *Cryomorphaceae bacterium* MED-G11 (4.16% of the FCB group) (Figure 2, red and dark blue nodes). Approximately 2% of classified reads belong to species *Nereida ignava*, compared to less than 0.04% from sample 2 and 3 it is the most prominent difference between the two locations (Figure 2, yellow node).

In the second sample four of the top-5 most abundant species belong to the same species: *Planktomarina temperata* (15.66% of Proteobacteria), *Flavobacteriales bacterium* UBA3446 (5.54% of the FCB group), *Flavobacteriales bacterium* UBA7358 (5.30% of the FCB group), *Flavobacteriales bacterium* UBA4585 (5.12% of the FCB group) and *Flavobacteriales bacterium* UBA7429 (4.41% of the FCB group) (Figure 3, red and dark blue nodes). Even though *Planktomarina temperata* reads are abundantly present in all three samples they are particularly enriched (15.66%) in this sample compared to 7.82% from the next year and 4.46% from France. Additionally, the presence of *Candidatus Pelagibacter ubique* is underrepresented in this sample, 1% of all classified reads belong to this species, compared to ~11% and 9% in sample 1 and 3, respectively (Figure 3, grey node).

Finally, the top-5 most abundant species from sample 3: *Candidatus Pelagibacter ubique* (9.24% of Proteobacteria), *Oceanospirillales bacterium* TMED91 (8.12% of Proteobacteria), *gamma proteobacterium* SCGC AAA168-P09 (8.08% of Proteobacteria), *Planktomarina temperata* (7,82% of Proteobacteria) and *gamma proteobacterium* SCGC AAA168-I18 (7.25% of Proteobacteria) (Figure 4, red and dark blue nodes). Interestingly, the species *gamma proteobacterium* are classified strain specific (Figure 4, dark blue nodes) as opposed to *Flavobacteriales bacterium* species from sample 2 and is less abundant in this sample (1.6%) compared to the year before (5.9%) (Figure 4, green node).

Comparison of OneCodex species classification between different locations and time interval and overall identified ranks

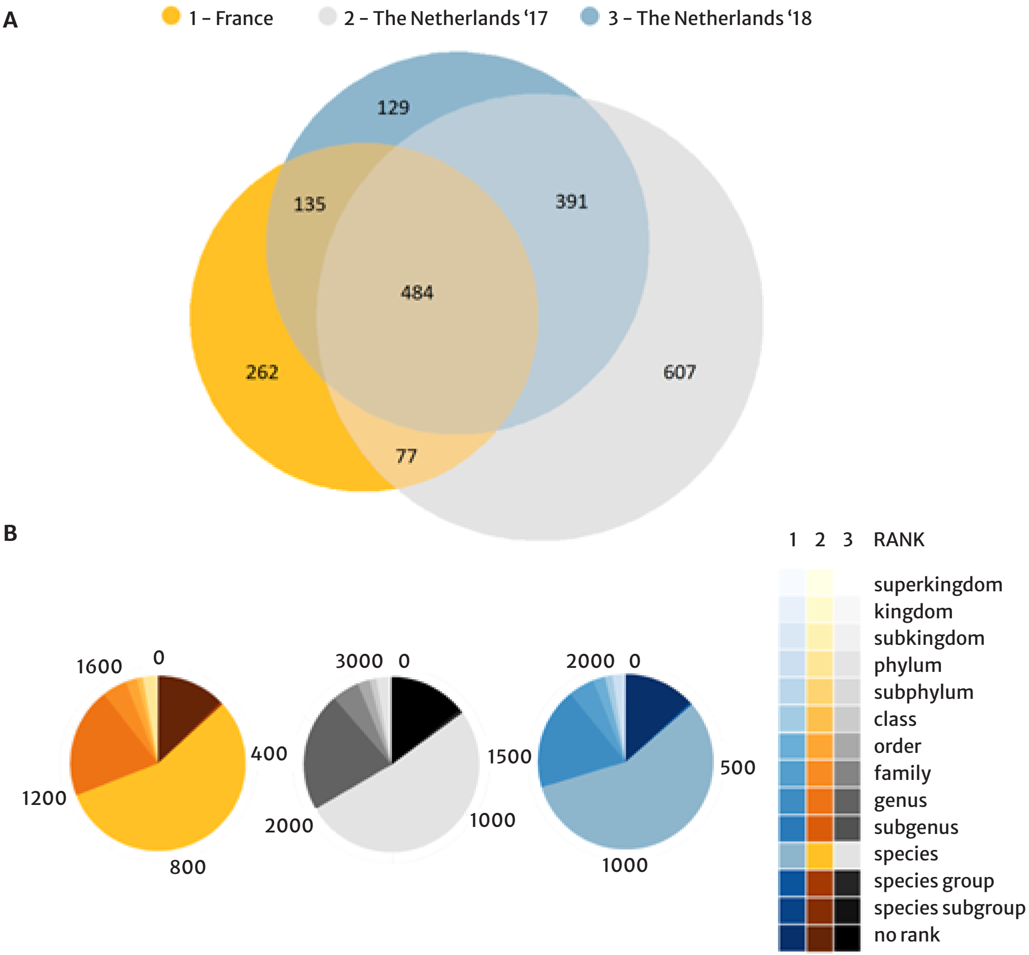

Figure 5 A) Venn diagram comparison of identified species by OneCodex, highlighting species that are time and space dependent and also microbes that are not.
B) Overall OneCodex classification ranks per dataset, the majority of classified reads have been linked to a species.

The taxonomic levels assigned by OneCodex range from kingdom down to species-specific. Reads that cannot be linked to a particular taxonomic level are labelled 'no rank'.
In total 1,750, 3,017 and 2,007 taxids are assigned to the data of sample 1, 2 and 3, respectively. More than half of the ranks that OneCodex was able to classify are assigned to species level (Figure 5 B) in all three samples.

Interestingly, at least 484 microbes are identified in all three samples (Figure 5 A). Some highlights include: 92 different *Flavobacteriaceae bacterium* and *Flavobacteriales bacterium* strains; 19 different *Candidatus Pelagibacter* strains; 18 *Pelagibacteraceae bacterium* and 6 SAR strains. This indicates that these communities are less time and location dependent compared to the 262 and 1,127 species that were found exclusively in France or Dutch areas, respectively. Furthermore, 607 and 129 species are exclusively observed in the Netherlands. As they exist at different times, they provide an initial impression of the time-dependent dynamics of these local communities. Finally, 135 and 77 species could be identified that are present at both locations, however only detectable at particular times. This could be an indication that even over large areas microbes are subject to time regulated dynamics.

## Metagenomics assembly on raw sequencing data and blast verification on the top-3 longest contigs

In an attempt to very OneCodex classification results as well as to assess the current metagenomics assemblers capabilities we subsequently assembled the three datasets separately. We have assembled our complex metagenomics datasets with Flye and retrieved 256, 1,735 and 968 contigs with mean coverage of 14x, 13x and 10x from samples 1, 2 and 3, respectively (Table 3). Coverage on contigs ranged up to 62, 89 and 107 for samples 1, 2 and 3, respectively, with a lower-bound of 3x coverage for all three assemblies. As expected, assembly statistics on sample 1 show the least optimal assembly results (lowest number of contigs, smallest mean and max contig lengths and smallest N50 values) given that the data volume of this sample was smallest combined with shortest average read lengths. Notably, although it has higher coverage, assembly results from sample 2 did not exceed results from sample 3. On the contrary, sample 3 resulted better average contig length, maximum contig length and N50 values compared to sample 2 (Table 3 and Table 4).

Table 3 Flye assembly statistics

| Assembly stats | France (1) | The Netherlands '17 (2) | The Netherlands '18 (3) |
|---|---|---|---|
| *contigs* | 256 | 1,735 | 968 |
| *length (bp)* | 8,678,102 | 107,863,873 | 94,117,952 |
| *min length (bp)* | 2,432 | 536 | 494 |
| *mean length (bp)* | 33,898 | 62,169 | 97,229 |
| *max length (bp)* | 219,363 | 1,098,797 | 1,648,106 |
| *N50* | 40,621 | 75,928 | 153,524 |

Impressively, Flye was able to reconstruct a full genome from our third sample: 75% of our 1.6 Mbp contig aligns with 80% identity to *Candidatus Thioglobus singularis* of which its complete genome is a single circular chromosome of 1.7 Mbp, with only 20x coverage on this particular contig (Table 4). The longest contig (219 Kbp) assembled from sample 1 represents a fragment of an entire genome and aligns with 88% identity to *Candidatus Pelagibacter ubique*, from which reads are most abundantly present in sample 1 (Table 4).

Even though OneCodex indicates that only 397 reads originate from *Candidatus Actinomarina*, Flye was able to reconstruct contigs that exceed the length of the currently available reference sequence. The second (141 Kbp) and third (137 Kbp) longest contigs aligned with 82% and 79% identity to the reference that is just 41 Kbp in size (Table 4). Similarly, Flye results in a top-3 longest contigs from sample 2 and 3 that align with high homology to the reference and all contigs exceed the length of the reference sequence (Table 4).

Table 4 **Blast alignment for top-3 longest contigs for sample 1, 2 and 3.** ID) identity number provided by Flye, Query len) the length of the contigs, Cont cov) data coverage for every contig, Best hits *) * criteria for best hit; largest query coverage with highest identity and published study, Query cov) how much of the contig covers the reference sequence, Aln ID) alignment identity between the reference and contig, Ref len) the length of the reference sequence the contig is aligned to.

| Sample | ID | Query len (Kbp) | Cont cov | Best hits * | Query cov (%) | Aln ID (%) | Ref len (Kbp) |
|---|---|---|---|---|---|---|---|
| 1 | 23 | 219 | 30 | *Candidatus Pelagibacter ubique* HTCC1062[17] | 88 | 78 | 1,308 |
| 1 | 227 | 141 | 16 | *Candidatus Actinomarina minuta*[16] | 24 | 82 | 41 |
| 1 | 130 | 137 | 13 | *Candidatus Actinomarina minuta*[16] | 16 | 79 | 36 |
| 2 | 190 | 1,098 | 24 | *Sphingobacterium* sp. EB080_L08E11[18] | 7 | 93 | 140 |
| 2 | 71 | 1,017 | 26 | *marine bacterium Betaproteobacterium*[19] | 10 | 94 | 44 |
| 2 | 8 | 967 | 27 | *marine bacterium Gammaproteobacterium*[20] | 4 | 80 | 61 |
| 3 | 58 | 1,648 | 20 | *Candidatus Thioglobus singularis*[28] | 75 | 80 | 1,714 |
| 3 | 376 | 1,283 | 7 | Uncultured *Flavobacteriia bacterium*[21] | 4 | 98 | 36 |
| 3 | 206 | 1,138 | 12 | *marine bacterium*<br>[AY458647] | 4 | 93 | 44 |

## Comparison of Flye assembly and raw sequencing data using OneCodex characterization

In order to verify if new species could be identified after assembly we have compared the OneCodex classifications using assembly results to the classification results based on raw sequencing data. Using the 256 contigs Flye was able to reconstruct OneCodex identified 41 species in total from sample 1 (Figure 6). Since reads that originate from *Flavobacteriaceae* and *Pelagibacteraceae* are represented in high abundance it is no surprise that detailed species-level classification for these two families appeared most effective, into 9 and 12 strains (out of the 41 classified species), respectively. OneCodex is able to identify 12 species only after assembly, these include 11 deferent *Pelagibacteraceae bacterium* strains and a SAR86 strain.

Although OneCodex was able to identify the most species using assembly results of sample 2, no prominent strain-specific enrichment was observed exclusively for assembly results in this sample. From the 209 species that are identified Flye favoured 5 species during assembly: *Alphaproteobacteria bacterium* (10 strains), *Euryarchaeota archaeon* (15 strains), *Flavobacteriaceae bacterium* (23 strains), *Flavobacteriales bacterium* (18 strains) and *Gammaproteobacteria bacterium* (19 strains).

Species diversification of assembly results from sample 3 appeared best for 14 different *Flavobacteriaceae bacterium* strains, 13 *gamma proteobacterium* strains, and 13 strains of *Gammaproteobacteria bacterium*. Notably, 6 *Pelagibacteraceae bacterium* strains could be identified using assembly results, that could not be classified based on raw sequencing data alone.

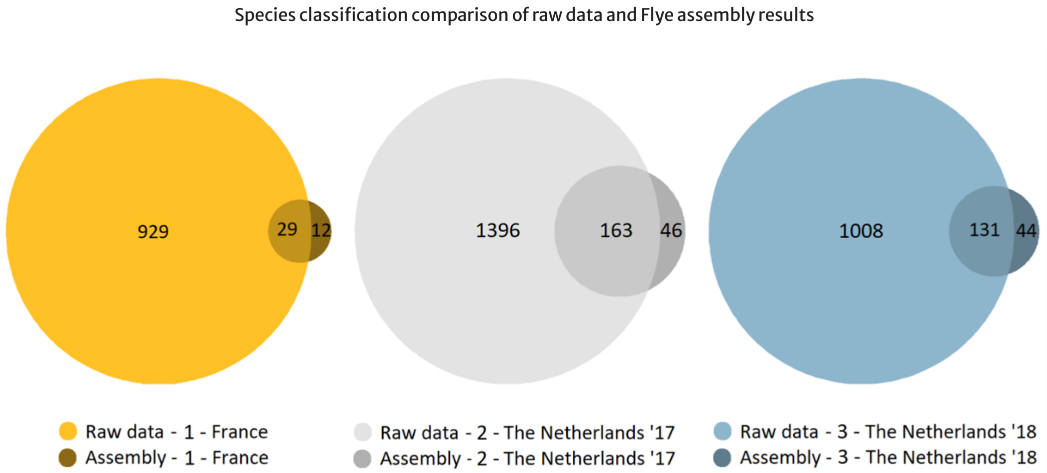

Figure 6 Species classification on sample 1,2 and 3.
Lighter shades indicate identified species on raw sequencing data, darker shades highlight species only identifiable after assembly.

### Data quality of unclassified reads and additional *in silico* PCR analysis

Poor read quality and relatively short read lengths could be a potential reason explaining why OneCodex was unable to classify taxids. Therefore, we investigated quality and length of unclassified reads (Figure 7). Although average lengths are shorter, and average quality values have a larger distribution, the differences are minimal compared to raw sequencing data (Figure 7). These statistics indicate that, in theory, the reads should provide OneCodex with sufficient information to resolve classifications. That OneCodex was not able to classify these reads even to the most general taxonomic levels (such as kingdom or phylum) adds to the notion that these reads originate from species that are novel.

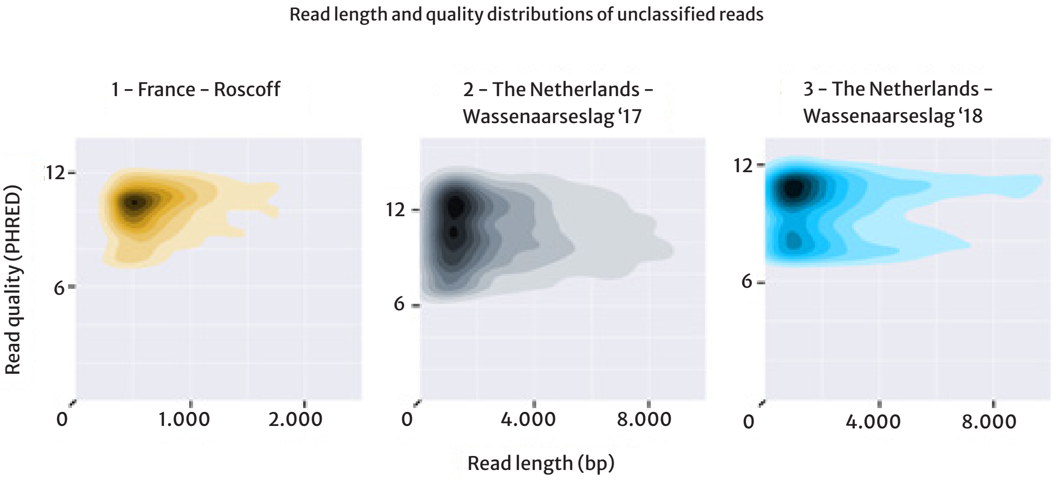

**Figure 7 Read length and quality distributions of data that OneCodex labels unclassified.** On average reads are shorter compared to raw sequencing data, however these lengths should still be sufficient to use for k-mer species characterization. Average quality distributions are very comparable to reads which OneCodex was able to classify species with.

The proportion of reads for which no classification could be assigned ranges between 38% and 69% compared to the raw sequencing data (Table 5) and provides a general impression on the amount of potentially novel microbes that thrive in these waters. Since OneCodex is particularly tailored to the identification of microbial DNA, unclassified reads potentially belong to non-microbial organisms. We therefore performed an additional round of *in silico* PCR analysis to inspect the presence of any remaining microbial 16S rRNA fragments. Interestingly, we found at least 10 more reads in sample 2 that have over 80% homology with our primers, showing that microbial content still exists within these unclassified reads (Table 5).

Table 5 Data statistics on reads for which OneCodex could not resolve any classification

| Stats | France<br>Unclassified (1) | The Netherlands '17<br>Unclassified (2) | The Netherlands '18<br>Unclassified (3) |
|---|---|---|---|
| *reads* | 172,843 | 908,744 | 86,653 |
| *bases* | 214,536,717 | 3,517,616,897 | 479,777,298 |
| *mean length* | 1,241 | 3,870 | 5,536 |
| *max length* | 33,295 | 155,979 | 132,486 |
| *% of original data* | 47 | 69 | 38 |
| *16S occurence* | 0 | 10 | 0 |

## Inspection of low complexity regions in unclassified reads using tandem repeat analysis

An additional circumstance that might explain why reads are left unclassified is the presence of low complexity regions such as repeat elements. These elements cause k-mers to contain the exact same genomic content making it impossible to assign them uniquely to specific taxa or even to a more general taxonomic level. We have analysed the presence of repeat elements with Tandem Repeat Finder[22] in raw sequencing data and compared these to repeat counts of the unclassified reads. In none of our samples did we observe an increased presence of repetitive elements, on the contrary, the repetitive element count is lowered in every case (Figure 8).

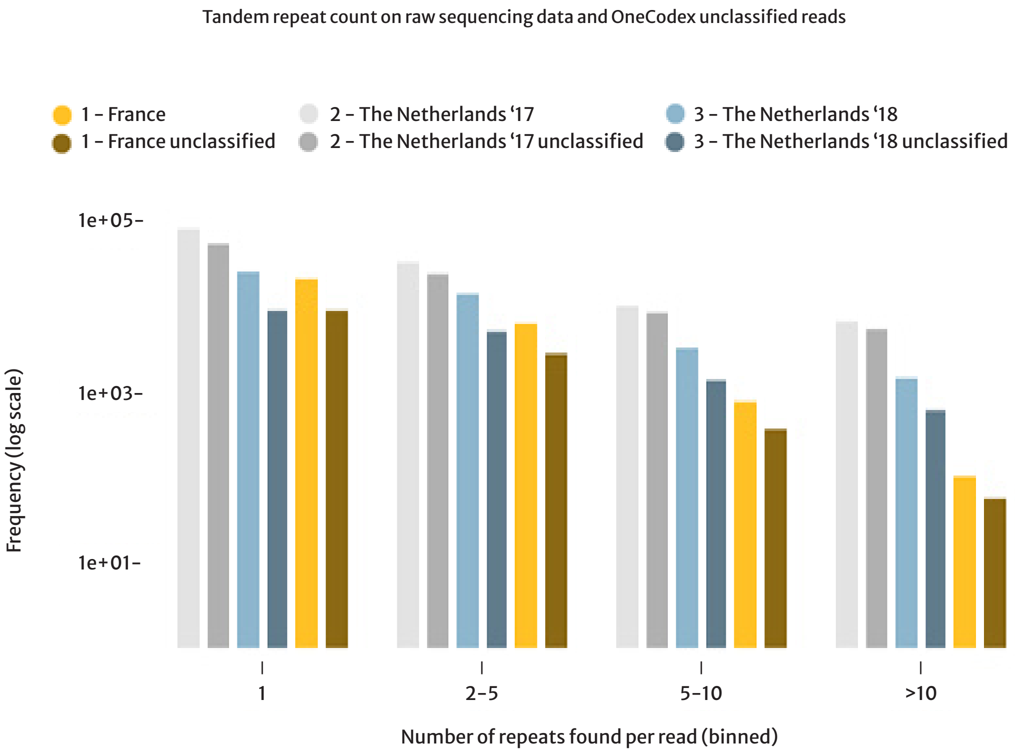

Figure 8 tandem repeat analysis, counts per read and comparison between raw sequencing data and unclassified data set for different locations and time. Repeat counts are represented in bins, the bins indicate the number of occurrences per read.

Taking together the data characteristics and the lack of both general taxonomical classification and highly abundant regions of low complexity suggest that these reads indeed originate from novel species. It highlights at least the absence of these species in currently publicly available OneCodex database, and provides a general glimpse of the amount of unknown species that comprise oceanic microbiomes.

## Sample collection and DNA isolation from salt water

Approximately 10 liter salt water of both locations was filtered through a double filter setup (Figure 9 A). 1.2 µm and 0.22 µm filters are used to remove eukaryotes and phages/ viruses from the samples, respectively (Figure 9 B). Water is passed through a 1.2 µm filter that aims to capture eukaryotic cells on top and is discarded, the remaining water is passed through a 0.22 µm filter during a second filtering round. The microbial material is captured on top of the filter, water that passed through this filter contains phage/viral material and is discarded afterwards. Material captured on the 0.22 µm filter represents the microbiome of our sample and was used for cell lysis.

Filtered biological sample and schematic respresentation of double filter setup

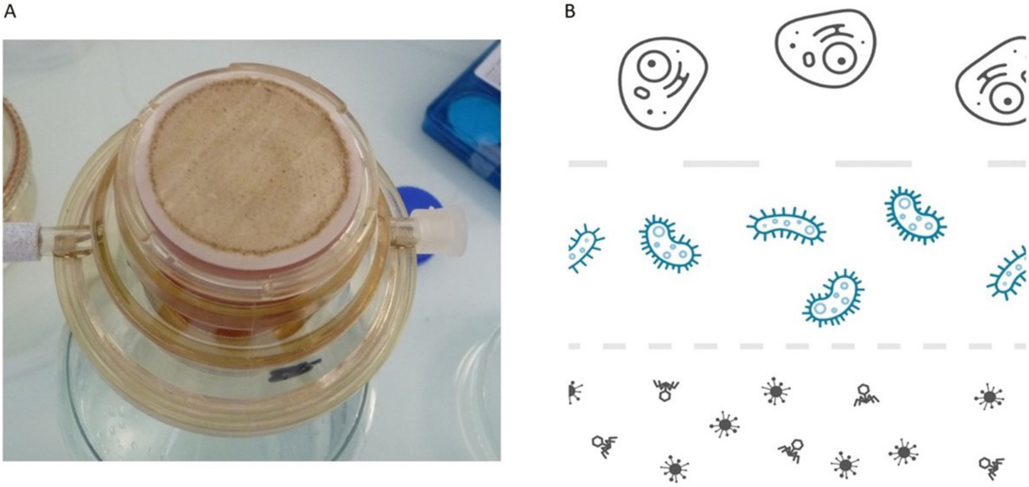

Figure 9 A) Filter setup; 0.22 µm containing biological material that represents the oceanic microbiome B) A schematic visualization of double filter setup. Discard eukaryotic cells during the first and viral/ phage content during the second filtering round.

To obtain high quality DNA we used the DNeasy PowerWater Kit (Qiagen), with minimal adjustments, according to the manufacturer's protocol. The largest adjustment was supplementing an enzyme set (Lysozyme, Mutanolysin and Lysostaphin) for a more extensive cell lysis. DNA from both North Sea samples was sequenced subsequent to DNA isolation, however we obtained a suboptimal yield from DNA isolation of sample 2 and amplified the isolation to meet the minimal input requirements for sequencing. Sample 1 was filtered through the double filter setup and temporarily stored at −20 °C and long term stored at −80 °C, DNA isolation and sequencing were performed after approximately 11 months of storage.

## DNA library preparation, sequencing, data quality control and statistics

We used R9.4 flow cells for sequencing all three seawater samples. Libraries were prepared using rapid kits (SQK-RAD004) available at that time according to the manufacturer's protocols (Oxford Nanopore Technologies, Oxford, UK). Data acquisition and base-calling were performed by MinKNOW (v19.06.8) controlling the MinION that sequenced the samples in 48-hours. Read-length and read-quality distributions were visualized using NanoPlot[23], and read counts, base counts and average read lengths were obtained using custom made scripts.

## Using *in silico* PCR analysis to verify microbial genomes

To highlight the presence of microbial genomes FastPCR[24] was used to perform *in silico* PCR analysis using primer pair sequences for identification of bacteria and archaea. FastPCR allows users to upload a set of primer sequences and reports, among others, positions and length of hits found on the input data. We used the currently 'best available' rRNA primer pair, primer 1 and 2 are 17 and 21 bp long, respectively, with a total amplicon size of 464 bp (primer 1: 5'-CCTACGG-GNGGCNGCAG-3', primer 2: 5'-GACTACNNGGGTATCTAATCC-3'). FastPCR verifies both forward and reverse primer sequences and due to the erroneous nature of our long read technology we have set a threshold of =>80% alignment identity to the primer sequences, with the exception that no errors may occur at the last position on the 3' end of the primer sequences.
Since OneCodex is primarily tailored to classification of microbial data we used FastPCR, in a similar fashion, to verify any remaining microbial content in the unclassified reads.

## K-mer based metagenome characterization of microbial sequences from seawater

OneCodex uses a k-mer based taxonomic classification algorithm to characterize microbial data. It uses a reference database containing 53,193, 27,020, 1,724, 1,756 and 168 bacterial, viral, fungal, archaeal and protozoan genomes, respectively. A default k-mer size of 31 bp is used to break up every read from the input data and compares them to a database that contains every k-mer that is uniquely linked to a taxonomic group. OneCodex classifies reads based on a set of k-mers that together uniquely identify taxonomic groups, single read hits are taken as the minimum threshold for identification in this study. OneCodex also provides reads for which no unique taxonomic classification could be found, we subtracted these reads from the initial input data using the command line interface (CLI) provided by OneCodex. We filter these reads using a project ID (provided by the web interface), the original dataset and set the taxonomic label to 0. For these reads we inspect the presence of microbial 16S rDNA and repetitive content in an attempt to explain the unresolved classification.

## Assembly of long read metagenomics samples using the Flye assembler

Flye[27] is currently one of the few *de novo* assembly pipelines that allows genomic reconstruction of complex metagenomics samples with coverage as low as 2x. We have downloaded the assembly software from the GitHub repository (v2.6), used the metagenome default settings and provided the raw sequencing data. For sample 1 and 3 we used all available raw sequencing data, for computational effectiveness we used half of the sample 2 data set. We have verified the top-3 longest contigs using BLAST alignment with high homology parameters and selected the best hits based on largest query coverage with highest identity and literature references.

## Repetitive content analysis for unclassified reads

To investigate the repetitive nature of reads that remained unclassified after OneCodex characterization we used Tandem Repeat Finder software (v4.09)[22], developed by Boston University, with default settings. The software locates repetitive patterns and reports their locations, sizes and copy numbers in a repeat table format. We have parsed both raw sequencing and unclassified data from sample 1, 2 and 3 to Tandem Repeat Finder and inspected the repeat occurrences on every read. With a custom-made script the frequencies of these occurrences on every read for every sample are summarized and expressed in 1, 2-5, 5-10 and >10 occurrences bins and plotted with R ggplot2[34].

## — Discussion

In this study, we have investigated the use of Nanopore sequencing for seawater metagenomics. Our main aims were to investigate the effectiveness of DNA isolation from samples directly obtained from the environment, optimize laboratory protocols for maximum sequencing results and evaluation of current metagenomics identification and assembly software. We used multiple isolation procedures, several different storage methods and subjected the data to a set of different analysis software. With only three ONT flow cells we were able to identify thousands of organisms, including bacteriophages, from which a large part at species level. It was possible to assemble genomes from environmental samples with Flye. In several cases this resulted in >1 Mbp contigs and in the particular case of a *Thioglobus singularis* species it even produced a near complete genome.

Although the enzyme cocktail used for cell lysis in our study was designed to break down cell walls for a wide range of bacteria there are potentially microbes that are immune to our lysis step. This might result in an underrepresentation of specific microbial communities compared to what truly thrives at these locations at that time. A possible solution, instead of lysing microbes with an enzyme set, would be to subject samples to mechanical lysis using silica beads or a combination of both. During experimental 12-hour sequencing runs (data not shown) we have observed that combining silica beads and enzymes during isolation yields significantly more sequencing data compared to isolation using only enzymes.

The double filter method separates eukaryotes and phages/viruses from bacteria in our sample. However, OneCodex still classifies a few hundred reads as either eukaryotic or viral. Eukaryota are particularly enriched for Dikarya, a subkingdom of fungi that are known to dominate the marine fungi fraction of environmental samples at European coastal regions[32]. These reads might have come from eukaryotic cells that are smaller than our largest filter (1.2 µm) or particles of these species that simply float around and were picked up by the smallest filter.

DNA molecules of our samples possibly suffered from fragmentation due to ice crystal formation during eleven months -20 and -80 oC storage. Additionally the yield of some sequencing runs is relatively low since biological material was dry frozen to the filter, making it more difficult to suspend the material during cell lysis. Under ideal circumstances DNA should be sequenced immediately after isolation circumventing DNA strand damage and loss of material.

The presence of viral DNA might be an indication that we have used too much water on a single filter, causing the accumulation of biological material to the point where the filter became saturated. A saturated filter might catch particles smaller than the smallest filter size and contaminate the isolation with material that would have otherwise passed through.
On the other hand, viral DNA could be present due to infection of microbes, which could be recognized by inspecting flanking regions of the read containing the viral DNA to contain microbe specific genes. Additionally viruses could enter the microbial metagenomics pool when they are present at the outside of bacteria and pass through the double filter setup via hitchhiking.

We initially performed *de novo* assembly in order to find out whether we could obtain longer contigs for particularly abundant species. Due to low coverage and the high diversity in our sample it is no surprise that this was possible for just a limited number of species. It is actually encouraging that with such diversity and limited sequence depth we could still identify more than thousand organisms at the species level. Moreover, metagenomics analysis is a relative newcomer in the field of genetics hence both laboratory protocols and analytical pipelines still need improvement to result higher accurate and more robust solutions for sample such as seawater.

We have performed an *in silico* PCR analysis to identify 16S RNA sequences in our raw sequence dataset. This showed that even under high error rate conditions reads contain enough homology to detect well conserved genes. This method could potentially be utilized to detect other genes in a similar fashion, for example genes that encode antibiotics biosynthesis.

While OneCodex was able to identify the diversity of a substantial amount of our samples, it could not resolve any classification for a large part of our data. The large k-mer size is most probably a crucial factor for unclassified data, due to the relatively low quality (approximately 10% error) of long-read data 10 bp would be a more suitable k-mer size. We confirmed that the data quality of these reads (both read length and quality distributions) are within acceptable bounds and observed no particular repetitive element enrichment compared to the reads that contributed to classifications.

Sequences that are representative of species that are currently unknown might explain the unclassified state of those reads, and are therefore valuable for contribution of a deeper understanding of the microbial marine fingerprint. Moreover, open access databases might not contain genomic information on particular microbes since obtaining genomes that are particularly large or come from non-culturable (non-culturable organisms are indicated with '*Candidatus*' labels) microbes remains a complicated task. For example, although available, protists are poorly represented in the OneCodex database, perhaps because their genomes are often extremely large (for dinoflagellates up to 270 Gbp[33]). Hence, microbes that are less thoroughly investigated might not have been included into the OneCodex genome selection. Since OneCodex is tailored to the identification of single cell organisms it probably will leave reads from multicellular organisms unclassified. Although no strong evidence was observed, lenient BLAST alignment of the top-3 longest reads of every sample did identify some small homologue regions with sequences from plant or algae in the NCBI database.

Despite the fact that these experiments are pilot studies we have observed promising results for both laboratory protocols and species identifications analysis. As described above, sample collection, DNA isolation and species identification is still hindered by both technical and biological difficulties. However our method provides a good impression on the elegance of our method that comes from its robustness and simplicity. We have performed equivalent experiments in student field practical assignments with similar marine samples, and students showed that even under more restricted conditions (12-hour sequencing runs) large biodiversity could still be detected. This indicates that the simplicity of our setup provides an ideal setting for student exercises, that will surely facilitate educational programs in genetics and bioinformatics.

## Data availability

Data submission in process and will be available at NCBI. The data is temporarily available at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA611514

## Declarations

Ethics approval and consent to participate – not applicable
Consent for publication – not applicable
Competing interests – authors declare no competing interest
Funding – No external funding
Authors' contributions – all authors contributed to the writing of the manuscript, ML performed bioinformatics analysis, AJGRT and ML performed lab experiments, ML and HPS wrote the first draft of the manuscript. CVH and HPS supervised the study.

## Acknowledgement

We would like to express our gratitude to OneCodex for answering questions on the available genome selection and the help with the CLI, and Future Genomics Technologies (Leiden) for the help with initial sequencing runs. All authors contributed equally to the manuscript.

## References

1. Zobell, C. E., Marine Microbiology, Chronica Botanica Co, Waltham, Mass., USA, 1946, p. 240.

2. Velankar, N. K., Bacteria isolated from seawater and marine mud off Mandapam (Gulf of Mannar and Palk Bay). Indian J. Fish., 1957, 4, 208–227.

3. Wood, E. J. F., Some aspects of marine microbiology. J. Mar. Biol. Assoc. India, 1959, 1, 26–32.

4. Marine microbial diversity. https://www.ncbi.nlm.nih.gov/pubmed/28586685

5. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. https://www.ncbi.nlm.nih.gov/pubmed/17355176/

6. Marine Rare Actinobacteria: Isolation, Characterization, and Strategies for Harnessing Bioactive Compounds (review) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5471306/

7. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. https://www.ncbi.nlm.nih.gov/pubmed/23959135

8. The Tara Pacific expedition-A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean https://www.ncbi.nlm.nih.gov/pubmed/31545807

9. Characterization of the microbial community diversity and composition of the coast of Lebanon: Potential for petroleum oil biodegradation https://www.ncbi.nlm.nih.gov/pubmed/31425842

10. Depth and location influence prokaryotic and eukaryotic microbial community structure in New Zealand fjords https://www.ncbi.nlm.nih.gov/pubmed/31377366

11. High-throughput sequencing and analysis of microbial communities in the mangrove swamps along the coast of Beibu Gulf in Guangxi, China. https://www.ncbi.nlm.nih.gov/pubmed/31253826

12. Environmental Genome Shotgun Sequencing of the Sargasso Sea https://www.ncbi.nlm.nih.gov/pubmed/15001713

13. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. https://www.ncbi.nlm.nih.gov/pubmed/17878949/

14. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system https://www.ncbi.nlm.nih.gov/pubmed/28584301

15. Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic. https://www.ncbi.nlm.nih.gov/pubmed/31344223

16. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3747508/

17. Genome Streamlining in a Cosmopolitan Oceanic Bacterium https://science.sciencemag.org/content/309/5738/1242.long

18. Time-series analyses of Monterey Bay coastal microbial picoplankton using a ‘genome proxy’ microarray. https://www.ncbi.nlm.nih.gov/pubmed/20695878

19. Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla https://www.ncbi.nlm.nih.gov/pubmed/17359257

20. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. https://www.ncbi.nlm.nih.gov/pubmed/14566056

21. Genomic content of uncultured Bacteroidetes from contrasting oceanic provinces in the North Atlantic Ocean. https://www.ncbi.nlm.nih.gov/pubmed/21895912

22. G. Benson,”Tandem repeats finder: a program to analyze DNA sequences” Nucleic Acids Research (1999) Vol. 27, No. 2, pp. 573-580.

23. Nanoplot Github webpage. https://github.com/wdecoster/NanoPlot, Accessed 30 January 2020.

24. Kalendar R, Khassenov B, Ramankulov Y, Samuilova O, Ivanov KI 2017. FastPCR: an *in silico* tool for fast primer and probe design and advanced sequence analysis. Genomics, 109: 312-319. DOI: 10.1016/j.ygeno.2017.05.005

25. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592464/

26. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification Samuel S Minot, Niklas Krumm, Nicholas B. GreenfieldPublished 2015 DOI:10.1101/027607

27. Kolmogorov M. et al. (2018) Assembly of Long Error-Prone Reads Using Repeat Graphs. https://doi.org/10.1093/bioinformatics/bty956.

28. Genome Sequence of “Candidatus Thioglobus singularis” Strain PS1, a Mixotroph from the SUP05 Clade of Marine Gammaproteobacteria. https://www.ncbi.nlm.nih.gov/pubmed/26494659

29. Abundant SAR11 viruses in the ocean. https://www.ncbi.nlm.nih.gov/pubmed/23407494

30. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. https://www.ncbi.nlm.nih.gov/pubmed/20861243

31. Complete Genome Sequence of the Fish Pathogen Flavobacterium columnare Strain C#2 https://www.ncbi.nlm.nih.gov/pubmed/27340080

32. Molecular diversity and distribution of marine fungi across 130 European environmental samples https://royalsocietypublishing.org/doi/10.1098/rspb.2015.2243

33. The exceptionally large genome of the harmful red tide dinoflagellate Cochlodinium polykrikoides Margalef (Dinophyceae): determination by flow cytometry, Algae 2016; 31(4): 373-378, DOI: https://doi.org/10.4490/algae.2016.31.12.6

34. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016

— Chapter 6

# Summary and discussion

## Summary introduction

In this thesis I highlight the applications of Oxford Nanopore Technologies (ONT) sequencing. This technique is a relatively new approach in the sequencing field, where nanopores are embedded in a membrane, DNA molecules are pulled through nanopores and an electrical current serving as the sequencing signal. This technique yields reads-lengths of >10Kbp and has no theoretical upper limit towards read-length. The positive impact on data quality due to improved chemistry is underlined, improved chemistry leads to less sequencing errors and a more homogeneous coverage over complex genomic architectures. Benefits for increased read-lengths are assessed for resolving fragmented genome assemblies that were previously based solely on short-read sequencing data. Furthermore, the assembly of a large genome using ONT data is described, indicating ONT is a suitable candidate for resolving extremely large genomes using sophisticated assembly approaches. And finally, the potential for on-site sequencing is evaluated. Exploiting simplicity, mobility and accuracy provided by this new technique.

The central hypothesis of this thesis is that Oxford Nanopore Technologies long-read sequencing can be valuable for established genomics applications, such as whole genome sequencing (chapters 2–4) and metagenomic characterization of microbial communities (chapter 5). Here, I reconsider this general proposition in light of the results of the preceding chapters. In addition, I discuss the prospects for emerging and future genomics applications based on the possibilities presented by ONT data.

## The quality of long- read sequencing and assemblies

ONT sequencing differs from traditional sequencing methods in the way that nucleotides are directly measured using electrical signals as opposed to synthetic copies or surrogate markers such as fluorescent labels. Multiple nucleotides (5-mers) occupy the pore shaft at the same time, hence it is the set of nucleotides that cause the electrical interference to determine the final profile. This profile signal needs to be, algorithmically, untangled to identify a single sequenced base. Therefore, basecalling algorithms yield the interpretable sequencing reads, and by improving basecalling algorithms sequencing data quality is subsequently increased even for previously sequenced projects[1]. ONT initially allowed 30 nucleotides per second to pass through the nanopore. The number of bases processed by a single pore was limited since basecalling algorithms struggled to differentiate nucleotides that passed through the pore too quickly, resulting in extremely low sequencing quality. Restricting the sequencing speed to 30 bases per second yielded ~70% accuracy. Currently ONT can process ~450 bases per second yielding reads >10Kbp with accuracy between ~90-99%. Chapter 2 highlights the effect of improved sequencing speed, basecalling and chemistry for a highly heterozygous yeast strain.

However, to generate accurate haplotypes for this genome additional sequence accuracy is required. From BUSCO analysis we observed genes that remained absent from our best assembly results. The difference in alignment hit, due to error introduction, is highlighted by the comparison of identified genes before and after error correction.

Where more genes are identified when sequence accuracy is increased. Compared to other studies, which use coverages ranging from 70x up to 1000x, our dataset has relatively low coverage. Hence slight coverage increase could aid in resolving any remaining assembly difficulties as well as increase sequence accuracy due to increased evidence for the error-correction procedure[2–4].

Since assembly algorithms struggle to define the ends of circular DNA, circularization for complex genomes remains a challenging task. In this study we have not investigated the architecture of mitochondrial DNA or circularization of plasmids. Hence it would be beneficial to subject the final assembly results to circularization software designed specifically for closing circular contigs from assemblers using long read data[2, 5].

In chapter 2 and 3 we have evaluated a multitude of assembly, consensus calling and correction tools, which have performed anywhere from mediocre to promising. Most assembly strategies are comparable and result in relatively small differences. Highlighting the origin of those small discrepancies and deciding upon the final assembly is tedious and a labor-intensive matter. The currently available tools leave room for user-friendly workflows, including base level and genome wide visualizations. These workflows should report progress at alignment, assembly, consensus, and correction level to facilitate decision making for downstream analysis. Off-the-shelf assembly workflows would increase the speed at which genome analysis is performed, and reliefs investigations for large sequencing datasets. The current gold-standard is performing multiple assembly strategies and continue in a result-based fashion. Analysis tools for small to medium-size genomes show comparable yet still slightly different assembly results causing comparison between analyses to be extremely difficult[7, 9].

*De novo* assembly results based on long-reads for small genomes show promising reconstructions. Assemblies for medium-large genome sizes of comparable quality, such as investigated in chapter 4, are increasingly publicly available. However, separated haplotypes of such organisms have yet to be uncovered since base-level quality has only recently become of sufficient quality to accurately phase chromosomal copies[10].

Despite increased capability towards evenly spread coverage, increased read-length and improvement towards low-complexity regions, for ultra large genomes additional development is required[6]. Sequencing ultra large genomes at routine level requires an additional developmental update particularly towards sequencing speed and cost. For instance, sequencing the genome of *Paris japonica*, a plant species with a genome size of unprecedented scale, estimated genome size ~150 Gbp for a single genome copy[8]. Sequencing a genome of this magnitude requires just under one hour on a fully loaded PromethION (that is 48 flow cells, each ~ $2.000 and utilizing ~ 2.500 pores at 450 bases per second) for a single genome copy. Hence, although feasible, sequencing at the required sequencing depth for such genomes still takes days and is very resource intensive. Evaluating the distinct ONT improvements towards read-length, read-accuracy and throughput, ONT is a pioneer entering the truly large-genome research area[6].

## The cost of genome sequencing

Evaluating the cost of genome sequencing using Moore's Law has made it clear that incredible amounts of sequencing data are going to be generated. These data volumes indicate the necessity of efficient downstream analysis software. Currently, sequencing data has become more affordable as opposed to costs for analyzing large datasets using computer clusters. The benefit of decreased cost and increased sequencing speed and throughput is lost when data analysis requires thousands of CPU hours on an expensive dedicated cluster. We therefore need to provide the scientific community with more sophisticated tools for processing large datasets, that are less computationally intensive, require less memory, are faster and more user friendly.

## Sequencing anything anywhere

Standard lab technicians are not experienced with command line tools and do not possess the skills to adequately adapt to alternative results. This clearly present gap could be bridged by using standardized metrics and formats, easily accessible free yet sophisticated software that is backed-up with logical visual representations.

In line with the skewed relation between generating and analyzing data is the size of sequencing machines, currently the smallest sequencing device is just the size of a large USB stick and provides mobility to allow infield sequencing, discussed in chapter 5. However, infield generated data needs to be processed by computer clusters or at least a high-end laptop with sufficient energy supply. Fully exploiting this mobility characteristic requires downscaled processing power and memory consumption.

## From amplicon to *in situ* metagenome sequencing and assembly

In chapter 5 we used metagenomics to identify the microbial diversity using ONT, which is a first step in understanding the oceans biocomplexity and ecology. However, to know which species thrive at which locations is only the start of understanding the ecology behind microbial diversity. To functionally assess microbial capabilities full genome assemblies are required, this would for example lead to increased understanding towards resistance mechanisms used by microbial communities to survive the harsh oceanic conditions or reveal the mechanistic property to interchange genomic content through plasmids. Additionally, it would highlight the diversification of species in a time and space fashion, enabling to monitor the health of oceans, seas and rivers that are the foundation of life on land.

To fully allow in-field monitoring of seawater, DNA isolation and library preparation methods need to be performed at location. In chapter 5 we have isolated the DNA under laboratory conditions. Although this procedure follows a very simple guideline, collecting high molecular weight DNA from marine organisms is particularly challenging due to excessive metabolites secretion that co-precipitates with DNA[11]. Hence optimization for high molecular weight DNA isolation regarding on site sequencing needs additional development towards speed and ease-of-use.

Additionally, isolated high molecular weight DNA requires library preparation to allow the sequencing device to bring molecules in proximity of sequencing pores and to read-out bases using an electrical current. Equipment for those preparations should meet desired requirements to be able for *in situ* use. Voltrax library preparation provides a potential solution and is able to prep isolated DNA in a matter of minutes, however, due to the lack of purification steps isolated high molecular weight DNA could be rather contaminated. Hence even with small and easy-to-use devices such as Voltrax, *in situ* DNA isolation and purification remains challenging[11].

Moreover, chemistry required for sequencing requires specific storage limitations; both flow cells and chemistry are temperature-sensitive and refrigerator capacity for in-field expeditions are usually inconsistent due to the lack of adequate power supply[12].

Finally, additional analysis is required to position identified species phylogenetically. Onecodex (used in chapter 5) is beneficial to place organisms quickly and easily into context of existing databases, easing time constraints and labor complexity. However, it lacks branch unity and cannot indicate the genetic distance between species and position them relative to each other. Furthermore, it only offers enhanced functionality using a paid license which adds to resource pressure and making it difficult for researchers to compare results. Previous studies show successful phylogenetic placement under remote conditions using JModelTest, hence this could be a potential candidate for downstream analysis on metagenome samples from seawater[13].

## The future of Oxford Nanopore Technologies sequencing and its applications

With the use of the current best flow cells and chemistry sequence accuracies of Q20 are achieved, translating to >99% read accuracy after basecalling. These methods allow molecules attached to the nanopore to be unzipped and both single strand copies are pulled through a pore reading-out the base sequence. Although sequencing both separated single strands was already introduced by Oxford Nanopore Technologies in ~2015, it was later replaced by single strand sequencing chemistry. However, chemistries to sequence both separated single strands have recently been released again by Oxford Nanopore Technologies. Here the information of both single strands is utilized to reduce basecalling errors by combining the sequencing signals. As the double stranded molecule found its way to the pore, one of the two strands is pulled through the pore, called the template strand. Subsequently unzipping the double stranded DNA leaves the 5' end of the complementary strand in proximity of the pore using a tether molecule attached to the membrane. As the sequencing reaches the end of the molecule, with some likelihood, the complement strand immediately follows the template strand through the same pore. From the output signal reads that transition one after the other with similar sequence lengths and complementary base composition are detected as pairs, referred to as a duplex pair.

Earlier basecalling methods either uses single strand signals or join signals from both template and complementary strands, called 'paired decoding'. On the one hand simplex basecalling (processing the signal of a single strand individually) is very fast however yields higher error rates. On the other hand, feeding both strands to a neural network basecalling algorithm decoding base pairs yield high accuracy sequences at the expense of resources and time. Decoding base pairs is computationally intensive, up to five times slower compared to simplex basecalling and therefore lacks scalability[14].

The novel quality increase for ‘stereo duplex basecalling’ finds its origin by feeding base information, quality scores and the sequencing signal for both template and complement strand to a ‘stereo’ basecaller. This basecalling method is simple, fast, and robust allowing for better scalability towards generating large amounts of data over a reasonable time frame while yielding Q30 sequencing reads. With read-quality approaching gold-standard sequencing platforms Oxford Nanopore Technologies appears a promising technique for analyses that require high accuracy on base pair level, such as SNP detection and haplotype identification, particularly for polyploid genomes.

Even though we observe an outpacing of Moore’s Law (Figure 9 - introduction) regarding sequencing cost, long read sequencing remains relatively expensive. Under more cost efficient conditions long-read sequencing is also a well-suited candidate for functional genomics analysis. The ability to prepare samples libraries without amplification circumvents the introduction of sequence-specific biases, where some molecules are underrepresented, and others excessively amplified allowing precise quantification. Long-read sequences can span full-length transcripts in a single read, hence avoiding complicated transcript assemblies, allowing simplified identification, and requiring fewer sequencing reads to identify the same number of genes compared to short read methods[15]. Furthermore, since full-length transcripts are recorded using single reads, they are exceptionally valuable for the characterization of structural variation such as isoforms. Isoforms can exhibit different functional properties and expression levels, and they are extremely difficult to determine using short reads. Additionally, structural variation is used across a broad spectrum of research areas, where it has shown significant importance to understand cancers in clinical settings all the way up to encoding target traits in agricultural studies. Structural variation spans, in many cases, Mbp stretches in the genome and are impossible to capture in a single read using gold-standard sequencing techniques. Hence those regions are sequenced in a fragmented fashion and reassembled to uncover the full structural variation using gold-standard techniques. This yields misassemblies and the absence of regions that are prone to amplification biases using other sequencing methods. Furthermore, since long reads provide increased alignment specificity the number of ambiguous alignments is significantly reduced, rescuing alignment regions that are lost using short read methods.

And finally, the sensitivity of sequencing signals and developments in artificial intelligence allows nanopore sequencing to detect modified bases. The epigenome is a complicated framework existing of a multitude of chemical compounds dictating DNA functionality. The higher order structure orchestrating the genome function comprises, among others, CpG methylation, nucleosome occupancy, chromatin accessibility, histone modifications and protein binding events that aid in proper segregation of chromosomes[16,17]. The most well-known epigenetics component is CpG methylation and is associated with suppressing gene transcription under hyper methylated promotor conditions or transcription activation for hypo- and hypermethylation of the promotor region and gene body, respectively. A gold-standard method to detect methylation is whole genome bisulfite sequencing, where unmethylated cytosines are replaced, at first using uracil and later by thymine nucleotides, revealing the methylation fingerprint. However, this method requires complicated bisulfite conversion steps, amplification and yields short reads. Hence this strategy is particularly difficult to apply for low complexity regions such as GC-islands. Oxford Nanopore Technologies methylation identification has been shown to perform with similar accuracy compared to gold-standard methods. In addition, it offers the benefit of increased read-length and the absence of amplification, allowing better alignments for low complexity regions, avoiding complicated laboratory procedures, and only utilizing the sensitive sequencing signal and adjusted basecalling algorithm[18]. Applications for functional genomics and epigenetics have proven their worth for specific scientific bottlenecks and have bridged knowledge gaps of areas left untouched by traditional technologies. The current cost perspective makes Oxford Nanopore Technologies specifically attractive for specialized cases, whether that is to identify genes surrounded by repetitive content, quantify splice variants with repetitive content, generating methylation fingerprints over long range epigenetic elements or to close assembly gaps for large and complex genomes. When Oxford Nanopore Technologies reaches a cost-effective ratio comparable to gold-stand methods it will find its true potential and will open a new era for standardized sequencing and data processing allowing the analysis of anything, by anyone, everywhere.

## To boldly go where no man has ever gone before

Suggested by the rapid read-length improvements it becomes more realistic to hypothesize that future sequencing will transform from a read-out of fragments method into a telomer-to-telomere sequencing fashion. Currently, the maximum read-lengths reported are >4 Mbp, compared to >10 Kbp during 2010 indicating it will not be long before end-to-end telomere sequencing is the gold standard. Sequencing entire chromosomes would bring significant benefits compared to current sequencing technologies, as it circumvents assembly for whole genome sequencing altogether. Downscaling computational load will relieve the scientific community of computationally intensive downstream analysis and will free scientist from dedicated computer clusters and command line tools.

Furthermore, sequencing speed is based on the number of nucleotides passing through the nanopore; to protect accuracy speeds are currently limited to 450 nucleotides per second. This sweet spot allows modern deep learning algorithms to determine the base sequence with up to Q30 accuracy. Increasing sequencing speed using those basecalling models would cause accuracy reduction as sequencing signals become too difficult to untangle. However, deep learning improvements resulting in more sophisticated neural network basecallers could increase sequencing speed up to a theoretically derived maximum of >$10^6$ nucleotides per second[19]. Exploiting the maximum sequencing speed could sequence a single human genome copy in just under two hours using a single pore. Such reduced computational pressure and increased sequencing speed will allow analysis of DNA content of any organism on a mediocre laptop in a matter of minutes, as opposed to a matter of days using dedicated and expensive computer clusters.

Additionally, standardized analysis workbenches should aid to reduce the time constraints even further, enabling scientists to navigate through the genomic content quickly and easily in a comprehensive, user-friendly, and visually appealing manner. Although read-lengths approach chromosome lengths, additional progress for chemistries must be made to, among others, avoid the entanglement of such long molecules during the isolation and unzipping of the double stranded DNA molecule.

Another potential application for future Oxford Nanopore Technologies that circumvents cell lysis to obtain high molecular weight DNA is the ability to sequence DNA/ RNA directly from the cell. Bringing the nucleus in proximity of the outer membrane and strategically incorporating a nanopore on both the nuclear envelop as well as on the outer membrane the nuclear interior could be connected to the sequencing pore. Using the intrinsic machinery that regulates proliferation to control entanglement and folding, DNA molecules can exit the nuclear envelope through the outer membrane into the sequencing pore. This would in turn bypass complicated entanglement of very large molecules and at the same time evade DNA molecule breakage that frequently occurs due to invasive laboratory procedures such as pipetting or mechanical lysis.

With one large leap of faith, in line with single cell nucleus sequencing, it might be possible to return the sequenced DNA or RNA through an additional feedback-pore. The unwinding of the DNA strands is then facilitated by proteins to collect and reposition proteins that are attached to the DNA strand. This allows the read-out of a single cell's entire genomic content without the need to sacrifice the sample. And would enable researchers to generate paired datasets that are statistically of incredible value avoiding biological variation on a cellular level.

## References

1. Lee J Kerkhof, (2021), Is Oxford Nanopore sequencing ready for analyzing complex microbiomes?, FEMS Microbiology Ecology, Volume 97, Issue 3, fiab001, https://doi.org/10.1093/femsec/fiab001

2. Giselle C. Martín-Hernández, et. al., (2021), Chromosome-level genome assembly and transcriptome-based annotation of the oleaginous yeast Rhodotorula toruloides CBS 14, Genomics, Volume 113, Issue 6, Pages 4022-4027, ISSN 0888-7543, https://doi.org/10.1016/j.ygeno.2021.10.006.

3. Min-Seung Jeon et al., (2023), Life Science Alliance, 6 (4) e202201744; DOI: 10.26508/lsa.202201744

4. Yury A Barbitoff et al., (2021), Chromosome-level genome assembly and structural variant analysis of two laboratory yeast strains from the Peterhof Genetic Collection lineage, G3 Genes|Genomes|Genetics, Volume 11, Issue 4, jkab029, https://doi.org/10.1093/g3journal/jkab029

5. Hunt, M. et al, (2015), Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16, 294. https://doi.org/10.1186/s13059-015-0849-0

6. Kathryn Dumschott et al., (2020), Oxford Nanopore sequencing: new opportunities for plant genomics?, Journal of Experimental Botany, Volume 71, Issue 18, Pages 5313–5322, https://doi.org/10.1093/jxb/eraa263

7. De Maio N et al., (2019), Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 5(9):e000294. doi: 10.1099/mgen.0.000294.

8. Jaume Pellicer et al., (2010), The largest eukaryotic genome of them all?, Botanical Journal of the Linnean Society, Volume 164, Issue 1, Pages 10–15, https://doi.org/10.1111/j.1095-8339.2010.01072.x

9. Segerman B (2020) The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front. Cell. Infect. Microbiol. 10:527102. doi: 10.3389/fcimb.2020.527102

10. Duan, H., Jones, A.W., Hewitt, T. et al., (2022), Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. Genome Biol 23, 84. https://doi.org/10.1186/s13059-022-02658-2

11. Sonia Boughattas et al., (2021), Whole genome sequencing of marine organisms by Oxford Nanopore Technologies: Assessment and optimization of HMW-DNA extraction protocols, Ecology and Evolution, https://doi.org/10.1002/ece3.8447

12. Aaron Pomerantz et al., (2018), Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building, GigaScience, Volume 7, Issue 4, giy033, https://doi.org/10.1093/gigascience/giy033

13. Darriba, D., Taboada, G., Doallo, R. et al. , (2012), jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772 https://doi.org/10.1038/nmeth.2109

14. Silvestre-Ryan, J., Holmes, I., (2021), Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. Genome Biol 22, 38. https://doi.org/10.1186/s13059-020-02255-1

15. Bayega, A., Oikonomopoulos, S., Gregoriou, ME. et al., (2021), Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci Rep 11, 7878. https://doi.org/10.1038/s41598-021-86753-7

16. Nicolas Altemose et al., (2022),Complete genomic and epigenetic maps of human centromeres. Science 376, eabl4178. DOI:10.1126/science.abl4178

17. Lee, I., Razaghi, R., Gilpatrick, T. et al., (2020), Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods 17, 1191–1199. https://doi.org/10.1038/s41592-020-01000-7

18. Mitchell R. Vollger et al., (2022), Segmental duplications and their variation in a complete human genome. Science376,eabj6965.DOI:10.1126/science.abj6965

19. Wang Y, Yang Q and Wang Z (2015) The evolution of nanopore sequencing. Front. Genet. 5:449. doi: 10.3389/fgene.2014.00449

— Addendum

# Nederlandse samenvatting - Dutch summary

## Introductie

In dit proefschrift focus ik op de toepassingen van Oxford Nanopore Technologies (ONT) sequencing. Deze techniek is een relatief nieuwe benadering in het sequencing-veld, waarbij nanoporiën zijn ingebed in een membraan, DNA-moleculen door nanoporiën worden getrokken en een elektrische stroom dient als het sequencing-signaal. Deze techniek levert sequenties (“reads”) van >10Kbp op en heeft theoretisch geen bovengrens voor de lengte van reads. De positieve impact op de datakwaliteit als gevolg van verbeterde chemie is uitgelicht, verbeterende chemie leidt tot minder sequentiefouten en een meer homogene verdeling van reads over complexe genomische architecturen. De voordelen van langere read lengtes zijn beoordeeld voor het oplossen van genoomassemblages die gefragmenteerd blijven met gebruik van uitsluitend korte-read-sequentiedata. Vervolgens is de assemblage van een groot genoom met ONT-data beschreven, wat laat zien dat ONT een geschikte kandidaat is voor het oplossen van extreem grote genomen met geavanceerde assemblagesoftware. En tot slot komt het potentieel van ONT sequencing naar voren voor in-het-veld sequencing, waarbij gebruik wordt gemaakt van de eenvoud, mobiliteit en de datakwaliteit die worden geboden door deze nieuwe techniek.

De centrale hypothese van dit proefschrift is dat Oxford Nanopore Technologies data waardevol kunnen zijn voor gevestigde genomics toepassingen, zoals volledige genoom sequencing (hoofdstukken 2–4) en het karakteriseren van metagenomen voor microbiële gemeenschappen (hoofdstuk 5). Hier evalueer ik deze algemene stelling in het kader van de beschreven resultaten van de voorgaande hoofdstukken. Daarnaast bespreek ik de vooruitzichten voor opkomende en toekomstige genomics-toepassingen op basis van de mogelijkheden die worden geboden door ONT data.

## De kwaliteit van long-read sequencing en assemblages

ONT-sequencing verschilt van traditionele sequencing-methoden doordat nucleotiden rechtstreeks worden gemeten met behulp van elektrische signalen in plaats van synthetische kopieën of markers zoals fluorescerende labels. Meerdere nucleotiden (5-mers) bezetten tegelijkertijd een porie, daarom is het de set van nucleotiden die de elektrische interferentie veroorzaken. Dit profielsignaal moet via algoritmes worden ontrafeld om een enkele base te identificeren. Het zijn dus de algoritmes die de uiteindelijke reads aanleveren en door deze algoritmes te verbeteren kan de kwaliteit van sequentie data zelfs verbeteren voor eerder geanalyseerde projecten[1]. ONT liet aanvankelijk 30 nucleotiden per seconde door de nanoporie passeren. De snelheid waarmee het aantal nucleotiden door een enkele porie werden gehaald werd gelimiteerd omdat algoritmes moeite hadden om nucleotiden te onderscheiden uit een set van nucleotiden die tegelijk de porie bezetten als deze te snel door de porie bewegen. Dit resulteerde in een bijzonder lage sequentiekwaliteit. Het beperken van de snelheid tot 30 basen per seconde leverde een nauwkeurigheid van ~70% op. Momenteel kan ONT ~450 basen per seconde doorlaten, wat reads oplevert van >10Kbp met een nauwkeurigheid tussen ~90-99%. Hoofdstuk 2 benadrukt het effect van verbeterde sequentiesnelheid, verbeterde algoritmes en chemie voor een zeer heterogene giststam.

Om echter nauwkeurige haplotypes voor dit genoom te genereren, is extra sequentiekwaliteit vereist. Uit een BUSCO-analyse bleek dat uit onze beste assemblageresultaten er nog steeds genen ongeïdentificeerd bleven. De impact van deze sequencing fouten wordt benadrukt door de vergelijking van geïdentificeerde genen vóór en na foutcorrectie. Waarbij meer genen worden geïdentificeerd wanneer de sequentienauwkeurigheid wordt verhoogd. Het volume van datasets word voor sequencing uitgedrukt in het aantal kopieën van het genoom ("coverage"). Vergeleken met andere studies, die coverage gebruiken variërend van 70x tot 1000x, heeft onze dataset relatief lage coverage. Daarom kan een toename van data helpen bij het oplossen van eventuele resterende assemblageproblemen, evenals het verhogen van de sequentiekwaliteit door meer bewijs te leveren voor de foutcorrectieprocedure[2-4].

Aangezien assemblagealgoritmen moeite hebben om de uiteinden van circulair DNA te definiëren is het assembleren van circulaire constructen voor complexe genomen een uitdagende taak. In deze studie hebben we de architectuur van mitochondriaal DNA of circulaire plasmiden niet onderzocht. Een logische volgende stap zou dus zijn om de assemblageresultaten te onderwerpen aan software dat specifiek is ontworpen voor het sluiten van circulaire contigs voor long-read data[2,5].

In hoofdstuk 2 en 3 hebben we een veelvoud aan assemblage-, consensus en correctietools geëvalueerd, die variëren van middelmatig tot veelbelovend. De meeste assemblagestrategieën zijn vergelijkbaar en resulteren in relatief kleine verschillen. Het benadrukken van de oorsprong van die kleine discrepanties en het beslissen over de uiteindelijke assemblage is een tijdrovende en arbeidsintensieve aangelegenheid. De momenteel beschikbare tools bieden ruimte voor verbetering van gebruiksvriendelijke workflows, inclusief visualisaties van base niveau tot genoomwijd. Deze workflows zouden voortgang moeten rapporteren op het niveau van alignment, assemblage, consensus en correctie om besluitvorming voor downstream analyse te faciliteren. Kant-en-klare assemblageworkflows zouden de snelheid waarmee genoomanalyse wordt uitgevoerd verhogen en onderzoeken verlichten voor grote sequentie-datasets. De huidige standaard is het uitvoeren van meerdere assemblagestrategieën en doorgaan op een resultaatgerichte manier. Analysetools voor genomen van klein tot middelgroot tonen vergelijkbare maar niet identieke assemblageresultaten waardoor vergelijking tussen analyses uiterst moeilijk is[7,9].

*De novo* assemblageresultaten op basis van long-read data voor kleine genomen laten veelbelovende reconstructies zien. Assemblages voor medium-grote genoomgroottes van vergelijkbare kwaliteit, zoals onderzocht in hoofdstuk 4, zijn steeds vaker openbaar beschikbaar. Echter, afzonderlijke haplotypen van dergelijke organismen moeten nog worden gepubliceerd, deze inhaalslag word nu pas gemaakt omdat de kwaliteit pas recentelijk van voldoende kwaliteit is geworden om chromosomale kopieën nauwkeurig te faseren[10].

Ondanks een verhoogde capaciteit om een gelijkmatige coverage te bereiken, langere reads te genereren en verbetering naar low-complexity regio's, zijn voor ultra-grote genomen aanvullende ontwikkeling vereist[6]. Het routinematig sequencen van ultra-grote genomen vereist een aanvullende ontwikkelingsupdate die met name gericht op de snelheid van sequencen en de kosten. Bijvoorbeeld, het sequencen van het genoom van *Paris japonica*, een plantensoort met een genoomgrootte van ongekende omvang, geschatte genoomgrootte ~150 Gbp voor een enkel genoomkopie[8]. Het sequencen van een genoom van deze omvang duurt iets minder dan een uur op een volledig geladen PromethION (dat wil zeggen 48 flowcellen, elk ~$2.000 en gebruikmakend van ~ 2.500 poriën bij 450 basen per seconde) voor een enkele genoomkopie. Daarom, hoewel haalbaar, duurt het sequencen op de vereiste sequentiediepte voor dergelijke genomen nog steeds dagen en is het zeer duur. Voor wat betreft de verbeteringen van read-lengte, read-kwaliteit en schaalbaarheid is ONT een pionier die het onderzoeksgebied van echt grote genomen mogelijk maakt[6].

## De kosten van genoomsequencing

Het evalueren van de kosten van genoomsequencing met behulp van de Wet van Moore heeft duidelijk gemaakt dat ongelooflijke hoeveelheden sequentiegegevens worden en zullen worden gegenereerd. Deze datavolumes geven de noodzaak aan van efficiënte software voor downstream analyse. Momenteel is sequencing-data betaalbaarder geworden in tegenstelling tot de kosten voor het analyseren van grote datasets met behulp van computerclusters.
Het voordeel van verminderde kosten, verhoogde sequencing-snelheid en schaalbaarheid gaat verloren wanneer gegevensanalyse duizenden CPU-uren vereist op dure toegewijde clusters.
We moeten daarom de wetenschappelijke gemeenschap voorzien van meer geavanceerde tools voor het verwerken van grote datasets, die minder rekenintensief zijn, minder geheugen vereisen, sneller zijn en gebruiksvriendelijker zijn.

## Alles en overal sequencen

Standaard laboratoriumtechnici hebben geen ervaring met commandline tools en beschikken niet over de vaardigheden om zich adequaat aan te passen aan alternatieve resultaten.
Dit duidelijk aanwezige hiaat kan worden overbrugd door gestandaardiseerde eenheden en formaten te gebruiken, gemakkelijk toegankelijke, gratis maar geavanceerde software die wordt ondersteund met logische visuele representaties.

Voor het idee alles overal kunnen sequencen is de omvang van sequencingmachines belangrijk, momenteel is het kleinste sequencingapparaat slechts zo groot als een grote USB-stick en biedt mobiliteit om sequencing in het veld mogelijk te maken, dit wordt besproken in hoofdstuk 5. Echter, veld gegenereerde gegevens moeten worden verwerkt door computerclusters of op zijn minst een high-end laptop met voldoende energievoorziening. Het volledig benutten van dit mobiliteitskenmerk vereist afgeschaalde verwerkingskracht, geheugen- en energieverbruik.

### Van amplicon tot *in situ* metagenoomsequencing en assemblage

In hoofdstuk 5 hebben we metagenomics gebruikt om de microbiële diversiteit te identificeren met behulp van ONT, wat een eerste stap is in het begrijpen van de biocomplexiteit en ecologie van de grote wateren. Echter, het bepalen welke soorten gedijen op welke locaties is slechts het begin van het begrijpen van de ecologie achter de microbiële diversiteit. Om deze diversiteit functioneel te beoordelen zijn volledige genoomassemblages nodig. Deze kennis kan bijvoorbeeld leiden tot een beter begrip van de resistentiemechanismen die door microbiële gemeenschappen worden gebruikt om de harde oceaansomstandigheden te overleven of om de mechanistische eigenschap te onthullen voor het uitwisselen van genetisch materiaal via plasmiden.

Bovendien zou het de diversificatie van soorten op een tijd- en ruimtelijke manier kunnen ontrafelen waardoor de gezondheid van oceanen, zeeën en rivieren die de basis van het leven op het land vormen, kan worden gevolgd. Om in-field monitoring van zeewater adequaat toe te passen moeten DNA-isolatie- en laboratoriummethoden ter plaatse worden uitgevoerd.

In hoofdstuk 5 hebben we het DNA onder laboratoriumomstandigheden geïsoleerd. Hoewel deze procedure een zeer eenvoudige richtlijn volgt, is het verzamelen van lang moleculair DNA van mariene organismen bijzonder uitdagend vanwege overmatige afscheiding van metabolieten die co-precipiteren met DNA[11]. Daarom moet optimalisatie voor isolatie van lang moleculair DNA met betrekking tot sequencing op locatie verder worden ontwikkelt zowel wat betreft sequencingsnelheid alswel wat betreft het gebruiksgemak. Apparatuur voor het voorbereidingen van DNA voor sequencing moet voldoen aan de gewenste eisen om *in situ* te kunnen worden ingezet. Voltrax laboratorium voorbereiding biedt een potentieel oplossing en is in staat om geïsoleerd DNA in een kwestie van minuten klaar te maken, echter, als gevolg van het gebrek aan zuiveringsstappen, zou geïsoleerd hoogmoleculair DNA nogal verontreinigd kunnen zijn.

Zelfs met kleine en gebruiksvriendelijke apparaten zoals Voltrax blijft *in situ* DNA-isolatie en -zuivering uitdagend[11]. Bovendien vereist de chemie die nodig is voor sequencing specifieke opslagbeperkingen; zowel flowcellen als chemie zijn temperatuurgevoelig en de koelkast-capaciteit voor veldexpedities is meestal onregelmatig vanwege het gebrek aan adequate stroomvoorziening[12]. Ten slotte is aanvullende analyse vereist om geïdentificeerde soorten fylogenetisch te positioneren. Onecodex (gebruikt in hoofdstuk 5) is gunstig om organismen snel en gemakkelijk in de context van bestaande databases te plaatsen, dit scheelt tijd en verlicht de arbeidscomplexiteit.

Aan het analyse portaal dat Onecodex biedt ontbreekt echter de phylogenetische afstand tussen soorten, welke naar boven gehaald kan worden door een tijdrovende methode zoals multiple sequence alignment. Bovendien biedt het alleen uitgebreide functionaliteit met een betaalde licenties waardoor kosten toenemen en het voor onderzoekers moeilijk maakt om resultaten te vergelijken. Eerdere studies tonen succesvolle fylogenetische plaatsing onder afgelegen omstandigheden met behulp van JModelTest, daarom zou dit een potentieel kandidaat kunnen zijn voor downstream-analyse van metagenoommonsters uit zeewater[13].

### De toekomst van Oxford Nanopore Technologies-sequencing en de toepassingen

Met het gebruik van de huidige beste flowcellen en chemie worden kwaliteiten van Q20 bereikt, wat zich vertaalt naar >99% nauwkeurigheid. Deze methoden maken het mogelijk om van de moleculen die door de nanoporie worden gehaald de basenvolgorde uit te lezen. Hoewel het sequencen van beide gescheiden enkelstrengs DNA al in ~2015 door Oxford Nanopore Technologies werd geïntroduceerd, werd het later vervangen door chemie van slechts één kopie uitleest. Echter, chemieën om beide gescheiden enkelstrengs DNA te sequencen zijn recentelijk opnieuw uitgebracht door Oxford Nanopore Technologies. Hier wordt de informatie van beide enkelstrengs DNA gebruikt om basecalling-fouten te verminderen door de sequentie-signalen te combineren. Zodra het dubbelstrengsmolecuul zijn weg naar de porie heeft gevonden, wordt één van de twee strengen door de porie getrokken, deze streng wordt de templatestreng genoemd. Vervolgens laat na ontvouwen van het dubbelstrengs-DNA het 5'-eind van de complementaire streng in de nabijheid van de porie achter met behulp van een bevestigings-molecuul dat aan het membraan is bevestigd. Naarmate de sequencing het einde van het molecuul bereikt, volgt met enige waarschijnlijkheid de complementaire streng onmiddellijk de templatestreng door dezelfde porie. Vanuit de sequencing signalen worden reads die na elkaar overgaan met vergelijkbare sequentielengtes en complementaire base-samenstelling gedetecteerd als paren, aangeduid als een duplexpaar.

Eerdere basecalling-methoden gebruiken ofwel signalen van enkelstrengs DNA of gecombineerde signalen van zowel template- als complementaire strengen, 'paired decoding' genoemd. Enerzijds is simplex basecalling (het verwerken van het signaal van een enkele streng individueel) zeer snel maar levert hogere foutpercentages op. Anderzijds, het voeden van beide strengen aan een neuraal netwerk basecalling-algoritme levert nauwkeurige sequenties op ten koste van middelen en tijd. Het decoderen van gecombineerde signalen is een rekenkracht intensief proces, tot wel vijf keer trager vergeleken met simplex basecalling en ontbreekt daardoor aan schaalbaarheid[14]. De noviteit van de kwaliteitsverbetering voor 'stereo duplex basecalling' vindt zijn oorsprong door base informatie, kwaliteitsscores en het sequentiesignaal voor zowel de template- als complementaire streng te voeden aan een 'stereo' basecaller. Deze basecalling-methode is eenvoudig, snel en robuust en maakt betere schaalbaarheid mogelijk om grote hoeveelheden gegevens te genereren over een redelijke tijdsperiode, welke Q30 kwaliteiten kan genereren. Met kwaliteit die de standaard sequencing-platforms benadert, lijkt Oxford Nanopore-technologie een veelbelovende techniek voor analyse die een hoge nauwkeurigheid op base niveau vereisen, zoals SNP-detectie en haplotype-identificatie, met name voor polyploïde genomen.

Hoewel we een overtreffing van de Wet van Moore (Figuur 9 - inleiding) zien wat betreft de kosten van sequencing in het algemeen, blijft long-read sequencing relatief duur. Onder meer kostenefficiënte omstandigheden is long-read sequencing ook een geschikte kandidaat voor functionele genomics-analyse. Het vermogen om samples voor te bereiden zonder amplificatie voorkomt de introductie van biases waarbij sommige moleculen ondervertegenwoordigd zijn en andere overmatig worden versterkt. Zonder deze biases kan nauwkeurige kwantificering mogelijk worden gemaakt. Long-read-sequenties kunnen volledige transcripten in één keer beslaan, waardoor ingewikkelde transcript-assemblages worden vermeden en vereenvoudigde identificatie mogelijk is. Hierdoor is er minder data nodig om hetzelfde aantal genen te identificeren in vergelijking met methoden voor short-read sequencing[15].
Bovendien zijn volledige transcripten die direct worden geregistreerd, uitzonderlijk waardevol voor de karakterisering van structurele variatie zoals isoformen. Isoformen kunnen verschillende functionele eigenschappen en expressieniveaus vertonen, en ze zijn uiterst moeilijk te bepalen met behulp van short-read sequencing. Bovendien wordt structurele variatie gebruikt over een breed spectrum van onderzoeksgebieden die lopen van het begrijpen van kankers in een klinische setting tot aan het coderen van commercieel aantrekkelijke eigenschappen voor de agrarische sector. Structurele variatie strekt zich in veel gevallen uit over Mbp-stukken in het genoom en is onmogelijk vast te leggen met een enkele read vanuit traditionele sequencing-technieken. Daarom worden die regio's, met traditionele data, sequentieel in stukjes gelezen en opnieuw samengesteld om de volledige structurele variatie te onthullen. Voor de standaard sequencing technieken leidt dit tot misassemblages en het ontbreken van regio's die vatbaar zijn voor amplificatie-biases. Bovendien, omdat long-reads een verhoogde aligneringspecificiteit bieden, wordt het aantal onduidelijke alignments aanzienlijk verminderd, in vergelijking met short-read sequencing data.

En tot slot, dankzij de gevoeligheid van sequentiesignalen en ontwikkelingen in kunstmatige intelligentie, kan nanopore-sequencing gemodificeerde basen detecteren. Het epigenoom is een ingewikkeld raamwerk bestaande uit een veelheid van chemische verbindingen die de functionaliteit van DNA dicteren. De hoog over structuur die de genomische functie orkestreert, omvat onder andere CpG-methylatie, nucleosoombezetting, chromatine-toegankelijkheid, histonmodificaties en proteïnebindende gebeurtenissen die helpen bij de juiste segregatie van chromosomen[16, 17]. Het meest bekende epigenetische component is CpG-methylatie en is geassocieerd met het onderdrukken van gen-transcriptie onder hypergemethyleerde promotoromstandigheden of transcriptieactivering voor hypo- en hypermethylering van het promotorgebied en een gen, respectievelijk. Een standaard methode om methylering te detecteren is whole genome bisulfite sequencing, waarbij ongemethyleerde cytosines worden vervangen, eerst met uracil en later door thymine nucleotiden, waardoor de methyleringsfingerprint wordt onthuld. Deze methode vereist echter ingewikkelde bisulfietconversiestappen, amplificatie en levert short-read data op. Daarom is deze strategie met name moeilijk toe te passen voor regio's met een lage complexiteit zoals GC-eilanden. Oxford Nanopore Technologies methyleringsidentificatie heeft aangetoond vergelijkbare nauwkeurigheid te behalen in vergelijking met standaard methoden. Bovendien bieden ze het voordeel van langere reads en de afwezigheid van amplificatie, wat betere alignments mogelijk maakt voor regio's met een lage complexiteit, het vermijd ingewikkelde laboratoriumprocedures, en heeft alleen het sequencing signaal nodig en een basecalling-algoritme[18].

Toepassingen voor functionele genomics en epigenetics hebben hun waarde bewezen voor specifieke wetenschappelijke knelpunten en hebben kennislacunes overbrugd van gebieden die onaangeroerd zijn gebleven door traditionele technologieën. Het huidige kostenperspectief maakt Oxford Nanopore Technologies specifiek aantrekkelijk voor gespecialiseerde gevallen, of dat nu is om genen te identificeren die omringd zijn door repetitieve sequenties, splice-varianten met repetitieve inhoud te kwantificeren, methyleringsfingerprints over lange reeksen epigenetische elementen te genereren of assemblage fragmenten te sluiten voor grote en complexe genomen. Wanneer Oxford Nanopore Technologies een kosteneffectieve verhouding bereikt die vergelijkbaar is met standaard methoden, zal het zijn ware potentieel vinden en zal het een nieuw tijdperk openen voor gestandaardiseerd sequencen, waardoor de analyse van "alles door iedereen, overal" mogelijk wordt.

## Moedig gaan waar niemand ooit geweest is

Zoals gesuggereerd wordt door de verbeteringen in read lengtes, wordt het realistischer om te hypothetiseren dat toekomstige sequencing zal transformeren van een methode voor het uitlezen van fragmenten naar een telomeer-tot-telomeer-sequencing-mode. Momenteel zijn de maximale leeslengtes die worden gerapporteerd >4 Mbp, in vergelijking met >10 Kbp in 2010, wat aangeeft dat het niet lang zal duren voordat telomeer-tot-telomeer-sequencing de standaard is. Het sequencen van hele chromosomen zou aanzienlijke voordelen met zich meebrengen in vergelijking met huidige sequencing-technologieën, omdat het de assemblage voor hele-genoom-sequencing volledig buitenspel zet. Het verkleinen van de computationele druk zal de wetenschappelijke gemeenschap verlichten van rekenintensieve downstream-analyses en zal wetenschappers bevrijden van toegewijde computerclusters en commandline software.

Bovendien is de sequencing snelheid gebaseerd op het aantal nucleotiden dat door de nanopore passeert, om de nauwkeurigheid te beschermen zijn de snelheden momenteel beperkt tot 450 nucleotiden per seconde. Deze snelheid maakt het mogelijk voor moderne deep learning algoritmen om de basenvolgorde met een nauwkeurigheid tot Q30 te bepalen. Het verhogen van de sequentiesnelheid met behulp van die basecalling-modellen zou echter leiden tot een vermindering van de nauwkeurigheid omdat sequentiesignalen te moeilijk worden om te achterhalen. Desalniettemin zouden verbeteringen in deep learning, resulterend in meer geavanceerde neurale netwerk basecallers, de sequentiesnelheid kunnen verhogen tot een theoretisch maximum van $>10^6$ nucleotiden per seconde[19]. Het benutten van de maximale sequencing snelheid zou een enkel kopie van het menselijk genoom in iets minder dan twee uur kunnen worden uitgelezen met behulp van een enkele porie. Een dergelijke verminderde computationele druk en verhoogde sequentiesnelheid zullen de analyse van DNA-inhoud van elk organisme op een middelmatige laptop in een kwestie van minuten mogelijk maken, in plaats van dagen met behulp van toegewijde en dure computerclusters.

Bovendien zouden gestandaardiseerde analyse-werkbanken moeten helpen om de tijdbeperkingen nog verder te verminderen, waardoor wetenschappers snel en gemakkelijk door de data kunnen navigeren op een uitgebreide, gebruiksvriendelijke en visueel aantrekkelijke manier. Hoewel read lengtes de lengtes van chromosomen benaderen, moet er aanvullende vooruitgang worden geboekt met laboratoriumtechnieken om, onder andere, verstrengeling of breken van dergelijke lange moleculen tijdens het isoleren en ontvouwen van het dubbelstrengs-DNA-molecuul te vermijden.

Een andere potentiële toepassing voor toekomstige Oxford Nanopore Technologies die cellysis omzeilt om lang moleculair DNA te verkrijgen, is het vermogen om DNA / RNA rechtstreeks uit de cel te sequencen. Door de kern in de nabijheid van het buitenmembraan te brengen en strategisch een nanopore op zowel de kern envelop als op het buitenmembraan te incorporeren, kan het binnenste van de kern worden verbonden met de sequentieporie.
Door gebruik te maken van het intrinsieke mechanisme dat proliferatie regelt om verstrengeling en vouwing te regelen, kunnen DNA-moleculen de kern envelop verlaten door het buiten-membraan in de sequentieporie. Dit zou op zijn beurt de ingewikkelde verstrengeling van zeer grote moleculen omzeilen en tegelijkertijd het breken van DNA-moleculen vermijden dat vaak voorkomt als gevolg van invasieve laboratoriumprocedures zoals pipetteren of mechanische lysis.

Met een beetje fantasie zou het zelfs mogelijk kunnen zijn om het uitgelezen DNA of RNA terug te voeren via een extra feedbackporie. Het afwikkelen van de DNA-strengen wordt dan gefaciliteerd door shaperone eiwitten die de losgekoppelde eiwitten verzameld en terug plaatst na het sequencen. Dit maakt de uitlezing van de volledige genomische inhoud van een enkele cel mogelijk zonder de noodzaak om de cel op te offeren. En zou onderzoekers in staat stellen om gepaarde datasets te genereren die statistisch enorm waardevol zijn, waarbij biologische variatie op cellulair niveau wordt vermeden.

## Literatuurlijst

1. Lee J Kerkhof, (2021), Is Oxford Nanopore sequencing ready for analyzing complex microbiomes?, FEMS Microbiology Ecology, Volume 97, Issue 3, fiab001, https://doi.org/10.1093/femsec/fiab001

2. Giselle C. Martín-Hernández, et. al., (2021), Chromosome-level genome assembly and transcriptome-based annotation of the oleaginous yeast Rhodotorula toruloides CBS 14, Genomics, Volume 113, Issue 6, Pages 4022-4027, ISSN 0888-7543, https://doi.org/10.1016/j.ygeno.2021.10.006.

3. Min-Seung Jeon et al., (2023), Life Science Alliance, 6 (4) e202201744; DOI: 10.26508/lsa.202201744

4. Yury A Barbitoff et al., (2021), Chromosome-level genome assembly and structural variant analysis of two laboratory yeast strains from the Peterhof Genetic Collection lineage, G3 Genes|Genomes|Genetics, Volume 11, Issue 4, jkab029, https://doi.org/10.1093/g3journal/jkab029

5. Hunt, M. et al, (2015), Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16, 294. https://doi.org/10.1186/s13059-015-0849-0

6. Kathryn Dumschott et al., (2020), Oxford Nanopore sequencing: new opportunities for plant genomics?, Journal of Experimental Botany, Volume 71, Issue 18, Pages 5313–5322, https://doi.org/10.1093/jxb/eraa263

7. De Maio N et al., (2019), Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 5(9):e000294. doi: 10.1099/mgen.0.000294.

8. Jaume Pellicer et al., (2010), The largest eukaryotic genome of them all?, Botanical Journal of the Linnean Society, Volume 164, Issue 1, Pages 10–15, https://doi.org/10.1111/j.1095-8339.2010.01072.x

9. Segerman B (2020) The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front. Cell. Infect. Microbiol. 10:527102. doi: 10.3389/fcimb.2020.527102

10. Duan, H., Jones, A.W., Hewitt, T. et al., (2022), Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. Genome Biol 23, 84. https://doi.org/10.1186/s13059-022-02658-2

11. Sonia Boughattas et al., (2021), Whole genome sequencing of marine organisms by Oxford Nanopore Technologies: Assessment and optimization of HMW-DNA extraction protocols, Ecology and Evolution, https://doi.org/10.1002/ece3.8447

12. Aaron Pomerantz et al., (2018), Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building, GigaScience, Volume 7, Issue 4, giy033, https://doi.org/10.1093/gigascience/giy033

13. Darriba, D., Taboada, G., Doallo, R. et al. , (2012), jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772 https://doi.org/10.1038/nmeth.2109

14. Silvestre-Ryan, J., Holmes, I., (2021), Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. Genome Biol 22, 38. https://doi.org/10.1186/s13059-020-02255-1

15.Bayega, A., Oikonomopoulos, S., Gregoriou, ME. et al., (2021), Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest. Sci Rep 11, 7878. https://doi.org/10.1038/s41598-021-86753-7

16. Nicolas Altemose et al., (2022),Complete genomic and epigenetic maps of human centromeres. Science 376, eabl4178. DOI:10.1126/science.abl4178

17. Lee, I., Razaghi, R., Gilpatrick, T. et al., (2020), Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods 17, 1191–1199. https://doi.org/10.1038/s41592-020-01000-7

18. Mitchell R. Vollger et al., (2022), Segmental duplications and their variation in a complete human genome. Science376,eabj6965.DOI:10.1126/science.abj6965

19. Wang Y, Yang Q and Wang Z (2015) The evolution of nanopore sequencing. Front. Genet. 5:449. doi: 10.3389/fgene.2014.00449

# Curriculum Vitae

Michael Liem was born on January 25th, 1987, in Alphen aan den Rijn, the Netherlands. He is the middle son for three children, his mother originating from the Netherlands and his father Indonesian Chinese. After finishing high school at Groene Hart Lyceum in Alphen aan den Rijn, he started his academic education with a BSc (in 2009) and MSc (in 2013) in Bioinformatics at Hogeschool Leiden and Leiden University, respectively. The subject of his MSc thesis was 'MAPK signaling classification in different cancers'. After finishing his MSc (in 2016), he started his Ph.D at the Institute of Biology Leiden (IBL) on the subject 'applications of nanopore sequencing' under supervision of Prof. dr. H.P. Spaink.

Currently, Michael is working as a data science teacher at Hogeschool Utrecht, where he educates Life Science students in the field of Bioinformatics. He is involved in a collaboration to incorporate a MSc track for Life Science students within the existing curriculum and works as researcher at the lectorate for Innovative Testing in Life Science and Chemistry.

# List of publications

Michael Liem, Hans J. Jansen, Ron P. Dirks, Christiaan V. Henkel, G. Paul H. van Heusden, Richard J.L.F. Lemmers, Trifa Omer, Shuai Shao, Peter J. Punt, Herman P. Spaink, (2018), *De novo* whole-genome assembly of a wild type yeast isolate using nanopore sequencing, F1000Research 2018, 6:618 doi: 10.12688/f1000research.11146.2

Hans J. Jansen, Michael Liem, Susanne A. Jong-Raadsen, Sylvie Dufour, Finn-Arne Weltzien, William Swinkels, Alex Koelewijn, Arjan P. Palstra, Bernd Pelster, Herman P. Spaink, Guido E. van den Thillart, Ron P. Dirks & Christiaan V. Henkel, (2017), Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads, Scientific Reports, 7: 7213, doi:10.1038/s41598-017-07650-6

Michael Liem, Tony Regensburg-Tuïnk, Christiaan V. Henkel, Hans Jansen and Herman Spaink, (2021), Microbial diversity characterization of seawater in a pilot study using Oxford Nanopore Technologies long-read sequencing, BMC Res Notes 14:42 doi.org/10.1186/s13104-021-05457-3