Near-complete genome assembly and annotation of a *Saccharomyces cerevisiae* wine strain utilizing third generation sequencing

Anu Bazarragchaa

Biology BSc, Third year

University of Debrecen

Faculty of Science and Technology

Department of Molecular Biotechnology and Microbiology

Supervisors: Dr. Valter Péter Pfliegler, Hanna Viktória Rácz

Debrecen

2021

Table of Contents

1. **Introduction** 3

1.1 *Saccharomyces cerevisiae* and its' role in alcohol industry 3

1.2 The *Saccharomyces* Genome 6

1.3 Whole genome sequencing (short- and long-read) 9

1.4 Whole genome assembly and annotation 10

2. **Aims & Objectives** 15

3. **Materials and Methods** 16

3.1. Strain maintenance 16

3.2. Genomic analysis 16

3.2.1. DNA extraction and purification 16

3.2.2. Library preparation and sequencing 16

3.2.3. *de novo* Genome assembly and annotation 17

4. Results 21

4.1 Sequencing quality statistics 21

4.2. Assembly and annotation pipeline comparison 21

4.2.1 Nuclear genome assemblies 22

4.2.2. Mitochondrial Genome Assembly 26

4.2.3. Annotation 27

5. Discussion 29

6. **Acknowledgements** 32

7. References 33

8. Appendix 41

1. Introduction

1.1 *Saccharomyces cerevisiae* and its' role in alcohol industry

History of yeast

One of the dozens of species extensively used in the food industry is *Saccharomyces cerevisiae*, commonly known as Baker's Yeast. The name *Saccharomyces* stands for sugar fungus, and *cerevisiae* stands for beer. Its use ranges from bakery products to different alcoholic beverages in our everyday life. The historical significance of yeast in post-agricultural civilization has been recognized. Various microorganisms, including yeasts, are presumed to have developed when people began cultivating and collecting significant amounts of crops (Legras et al., 2007).

Wine's history began thousands of years ago– as the earliest instances on record have been found in the Caucasus, Mainland China, and excavated on Egyptian papyri. It is assumed that winemaking came after wild vine (*Vitis vinifera* spp. *sylvestris*) domestication. *V. vinifera* was originally found in the Caucasus and Turkey. Although the origins of wine are disputed, it is most likely that the stockpiling of grapes in big containers in the Caucasus and its surrounding areas caused the fruit to rot, providing an ideal environment for yeasts to thrive (Johnson, 2019).

After the Ancient Greeks planted the seeds of winemaking in the north and southern Gaul, Rome's great vintages and the culture that followed laid down the foundation of what is now the culture of modern viticulture in Europe. The spread of winemaking is said to be a direct result of Greek ships traveling and selling wines at international ports as a commercial product (Johnson, 2019).

When the Romans started using earthenware amphorae and the Gauls (or Modern Germany) used wooden barrels to store wine, diversity in wine production techniques may have burgeoned. This culture quickly diminished during the Dark Ages, where the Church was the sole motivator for a better quality of wine. Vineyards were refined, spread throughout Western Europe until the 19th century due to their long shelf life. Several historical and biological factors affected the industry over the centuries, where the quality of wine dwindled (Johnson, 2019).

In the coming years, Pasteur's discovery and the next wave of industrialization changed the industry drastically. The discovery of fermentation's microbiological and physiological background, a process where microorganisms convert sugars into alcohol in an oxygen-depleted

environment, enabled oenologists and scientists to focus on yeasts more than ever (Johnson, 2019).

This species and its unique characteristics were largely unrecognized until the 20th century. Emil Christian Hansen of Carlsberg Laboratory isolated the first pure yeast culture from wine musts in the 1890s. Although wine quality and profile are affected by various factors during its manufacturing, the starter yeast added to wine must is one of the key attributes that help distillers produce their desired product. This is one of the main reasons why the quality of commercial wine yeasts sold on the market is crucial to distillers and winemakers. (Walther, 2014)

The yeast as an organism

The yeast is a single-celled organism with an ovoid-round shape and a diameter of 5–10µm. The cell is larger than most bacteria cells (0.4–3 µm) despite being smaller than normal human cells (10–100 µm). Yeasts may be found on the surface of fruits and flowers, but it is unknown where this species dwells (Goddard, 2015).

S. cerevisiae reproduces via budding, an asexual reproduction method in which a bud, also known as a daughter cell, emerges after cell division at the budding site. DNA replication starts off the budding process. After the large bud gradually separates from the mother cell, it completes one cell cycle. One cell cycle spans approximately 100 minutes and dies after 26 cycles. Wine yeast prefers to be in an environment of 30–35 °C, and its growth rate increases logarithmically. As its ethanol output increases, the cells start to die – a distinguishing characteristic for different strains that gradually determine the wine's alcohol content (Matallana, 2017).

Yeast is traditionally used in wine production due to its expansive capacity to carry on the fermentation process. This species uses the pre-existing glucose in its surroundings to generate ethanol and CO_2, which gives wine its alcohol concentration and various flavor profiles.

Fermentation

Yeasts play an important role in alcohol manufacturing due to their ability to produce ethanol in large amounts. The wine must's high acid and sugar content, low pH, and the growing presence of ethanol prevents other bacteria from competing with the yeast, allowing it to thrive

(Fleet G.H. 1998). Wine yeast, in particular, comes from cellars and is described to be domesticated to some extent several times (Fay & Benavides, 2005; Legras et al., 2007; Peter et al., 2018).

The organism is also known for its consistent production of byproducts under unfavorable conditions, such as osmolarity difficulties and low pH (Parapouli, 2020). This is the main cause of secondary metabolite production (Polsinelli et al., 1996), and selection pressures imposed by the environment's chemical composition, such as ethanol (Querol & Ramón, 1996).

Fermentation occurs when the presence of oxygen is depleted – while the electron transport chain is unable to work properly, the cell triggers a pathway that produces carbon dioxide and ethanol after glycolysis. Yeast cells digest sugars such as sucrose, fructose and glucose, maltose, and trehalose. Its byproducts: ethanol, and CO_2, gradually mix with the preexisting environment. The nutritional requirements of yeast are straightforward: a carbon and energy source sugar, and a nitrogen source such as urea, or ammonium ions, along with micronutrients.

In short, the primary role of yeast is to convert sugars into ethanol in wine production. In the presence of oxygen, yeast metabolism of glucose creates a considerable quantity of energy as well as a variety of intermediates that the cell requires to develop. In the absence of oxygen, the cell will perform fermentation – producing ethanol as a waste product. In most species of Saccharomyces, the Crabtree effect is observed: respiration is repressed, and in turn, the organism yields higher amounts of ATP in O_2 presence due to accelerated glycolysis.

Yeasts' importance in oenology

Strains of *S. cerevisiae* ferment the material, which gives the beverage flavor by using maltose and maltotriose (Barnett 1992). It also produces esters, numerous alcohols, aldehydes, and sulfur compounds in varying volatility which plays a crucial role in the product's distinctive taste. Modern breweries tend to use a range of *S. cerevisiae* strains to produce different products. As novel strains form due to their rapidly evolving nature, strain purity is monitored and tested using molecular methods for quality assurance in the industry (Johnston, 1990; Schofield et al., 1995, Gutteridge & Priest, 1996; Wightman et al., 1996).

Although *S. cerevisiae* may dominate the fermentation processes in wine production, studies suggest that the genomic data samples are not always purely made up of *S. cerevisiae*.

They may contain portions of genetic material from other yeast species and bacterial species (Langdon, 2019). This implies the coexistence of several strains of yeasts in extreme conditions together with a possibility for genetic exchange.

1.2 The *Saccharomyces* Genome

The yeast genome

The yeast is a model organism among many others in microbiology due to its speed of reproduction, genetic variance, and size of the genome. This easily translates to bigger research fields such as oncology and bioinformatics. The *Saccharomyces cerevisiae* genome was the first eukaryote scientists sequenced – now considered a huge milestone for population genomics and genetics. Its genotype-phenotype map was explored as a first step towards modern genetics (Liti, 2009; Peter et al., 2018).

This organism is diplontic in nature with a high likelihood of clonal reproduction. It is also homothallic, and the way of its genome renewal can be done by the regeneration of a diploid from a haploid (Mortimer et al., 1994). Aneuploidy is often perceived in wine yeasts (Bakalinsky & Snow, 1990). Due to its unicellular nature and rapid pace of reproduction, it is considered a valuable tool for eukaryotic genome research.

During its evolutionary history, *S. cerevisiae* faced several evolutionary pressures that produced different yeast strains over centuries with their respective set of uniquely evolved genomes. These phenotypes are now used in phenotypic analysis due to their highly specialized characteristics and specific applications (Legras et al., 2007; Fay, 2013; Peter et al., 2018).

The Yeast Genome Database

We used the S288c reference genome for the genomic assembly in this study. The S288c reference genome of *S. cerevisiae* is available on multiple sources, namely the NCBI's PubMed portal and the official *Saccharomyces* Genome Database, as an open-source material. It was first published in a May 1997 issue of Nature as the "Yeast genome directory" (Zagulski et al., 1998). The yeast genome consists of 16 chromosomes, around 6000 genes which amount to approximately 12.1 million base pairs. It shares 23% of genes homologous with the human genome, which makes it a useful model organism (Jiang, 2008).

Genomics of wine yeasts

Ecology-wise, wine strains do not have much variation in their genetic materials, as it was found that certain yeasts belonging to different ecological niches have different adaptations regarding their population structures (Schacherer, 2009; Liti, 2009). Wine yeasts are often put

through extreme conditions for the purpose of getting specific flavor notes or consistency. The domain of yeasts that are used in wines does not differ much due to such aggressive selection (Borneman, 2016). Most commercial strains on the market are largely similar to one another, with subtle differences.

In such extreme conditions, high amounts of ethanol affect selectively on mitochondrial DNA in one investigation. It caused variations in their mitochondrial DNA. The size and quantity of chromosomes varies depending on the form of reproduction, however, in meiosis, it had a greater rate of recurrence. This in turn, amplifies certain regions, thus increasing the likelihood of polymorphism and genomic variability (Benitex, 2002).

Due to its rapid speed of evolution, yeast still holds to be the main interest of researchers in genetics and evolution. Equipped with a rapid speed of reproduction and a unique subset of ways to reproduce, genetic variety in yeasts are under investigation.

These variations occur in polymorphisms such as single nucleotide polymorphisms (SNP). Available data suggests most papers dealing with genetic variation are more attuned to single nucleotide variant (SNV) levels. SNV and SNP are terms used interchangeably. On a bigger scale, structural variants (SV) such as transpositions and duplications are not explored thoroughly due to technical problems while working with short-read data. This is a problem because large-scale SVs contribute the most to speciation and disease in many organisms (Rieseberg, L.H., 2001).

Potentially adaptive copy number variants (CNV) in wine yeast isolates include regions containing genes encoding different transporters. Dehydrogenases and genes involved in the metabolism and efflux of toxic substances used in winemaking, such as copper and sulfite, are also present (Borneman A.R., 2011; Dunn B., 2005; Carreto L., 2008). Some strains also contain CNVs in individual pangenomic open reading frames. (Peter et al., 2018).

Gene copy numbers involved in the absorption and digestion of maltose, for example, are frequently increased in beer strains. This is attributed to the maltose-containing wort used in beer production, but within wine strains, they are frequently lost. Maltose is seldom found in grape musts (Gallone, 2016; Gonçalves, 2016).

There are also cases wherein adaptation is mediated by large structural variation rather than gene duplication. Adaptive reciprocal translocations between chromosomes VIII and XVI and chromosomes XV and XVI are common in wine yeasts. In a sulfite-containing media, both

translocations result in a shorter lag period. Sulfate is a typical preservative used to keep wines from oxidizing (Treu, 2014; Goto-Yamamoto, 1998; Pérez-Ortın, 2002; Yuasa, 2004; Zimmer, 2014).

Certain factors affect the likelihood of genetic variance. TEs are abundant in genomes, accounting for up to 90% of certain wheat genomes. Some yeast strains do not contain TEs, whereas others do to some level. Because TEs accumulate in genomes, they have a direct impact on genome structure, plasticity, genetic diversity, and evolution. They also use epigenetic processes to influence gene expression, structure, and function (Slotkin et al., 2007; Dominguez et al., 2018).

Telomeric core X and Y elements have been found to directly affect the Telomere Position Effect (TPE) in one study, which hints at the possibility of gene silencing and gene expression through these values (Power et al., 2011). Ty elements are unstable subtelomeric regions that cause diversity in yeast strains. These sequences are homologous to various regions in the genome (Pryde et al., 1995; Brown et al, 2010; Nijkamp et al., 2012).

Genes located near telomeres are transcriptionally suppressed in *S. cerevisiae*. This is called the telomere position effect (TPE). Although all chromosomal ends have the same telomeric DNA sequence, the subtelomeric elements (STEs) differ. There are two types of STEs in yeasts, called X and Y prime (stylized as Y') elements. X is a heterogeneous region ranging in size from 300bp to 3kb. All telomeres contain a variation known as "core-X". The core X contains replicator sequences, as well as a consensus sequence that aids in the binding of transcription factors and subunits. The core X element can improve gene silencing in certain conditions (Louis, 1995).

The Y′ element, in contrast to the ubiquitous X element, is only found in one-half to two-thirds of yeast telomeres. Binding sites can also be found in the Y′ element. When present, Y′ is found in up to four tandem copies and is distal to X (Chan & Tye, 1983). Even though there are only two types of STEs, the X and Y′ components are different enough that each of the 32 yeast telomeres may be considered unique (Mondoux, 2007).

Transposable elements (TE) are DNA segments that may travel throughout the genome, or transpose. Non-allelic homologous recombination (NAHR), the mispairing of two sections of extremely similar DNA sequences, such as comparable TEs, causes deletions, segmental

duplications, and inversions as a consequence of post-insertion effects (Bailey, 2003; Cordaux, 2009; Hancks, 2012).

1.3 Whole genome sequencing (short- and long-read)

Next generation sequencing (NGS) is a widespread technology used in genomics to investigate genetic variation (Nakagawa, 2018). Whole-genome sequencing (WGS) is a part of this technology, where an organism's entire genome is sequenced to find the exact order of bases. Through WGS, it is possible to link genetic information with corresponding phenotypes, which enables scientists to openly interpret the data and explore its genetic variation (Mardis, 2008; Ng & Kirkness, 2010).

WGS typically starts with DNA purification, followed by library preparation. Traditionally, short-read Illumina sequencing is performed in laboratories due to its low-cost and efficiency (Pollard, 2018). However, short read data has little access to data of GC-rich regions, so it may not be able to display complex regions of the genome correctly. Additionally, the repetitive, paralogous regions or large-scale structural variations (SVs) are not always mapped correctly (Horak, 2016; Alioto *et al*., 2016; Treangen & Salzberg, 2012). Long-read sequencing overcomes these pitfalls by lengthening the cut samples (Jain *et al., 2018;* Vollger *et al., 2020*), which in turn, reduces the amount of disorganized reads and increases SV detection sensitivity (Ebbert *et al.,* 2019).

Long-read sequencing generates longer reads which range from 10,000 up to 100,000 base pairs compared to short reads, which span up to a few hundred base pairs maximum. It spans the high complexity and repetitive regions to create accurate assemblies. Long-read sequencing technology captures the complex, paralogous genomic regions and detects structural variants such as gene loss, segmental duplications, and fusions (Hosomichi, 2013; Wang, 2016).

The Oxford Nanopore Technologies' (ONT) MinION is based on the recorded changes of electrical signals through a nanopore with enzymes attached. As nucleotides of a DNA molecule travel through a nanopore, this technique monitors changes in ion flow and generates a file in FAST5 format. The signal is amplified and measured to determine the bases that passed through. The pore is able to hold up to 5 bases at a time, so overlapping k-mers may sometimes interfere with data. All potential subsequences from a read obtained from DNA Sequencing are referred to

as k-mers. The Oxford Nanopore MinION does this on a minor scale with a competent error rate that is usually context specific (Pollard, 2018).

1.4 Whole genome assembly and annotation

Long-read (and short-read) based *de novo* whole genome assembly (WGA) brings together unique sections of the genome, and then characterizes the remaining difficult-to-assemble areas based on the evidence. It puts the nucleotide sequences in the correct order. This helps to minimize mistakes in further individual assemblies while creating the most accurate draft of a genome's overall structure (Brendan L., 2003). Contig assemblies are created from sequence overlaps between reads. Repetitive regions may be incorrectly assembled with earlier technologies (Samuel L., 2010). Because sequence-read lengths are substantially shorter than most genes in no particular order after sequencing, assembly is an important aspect of this study. After de novo assembly, to further enhance the quality of the assembly, long- and/or short-read based polishing is used (Vaser, 2017). To accurately position the contigs at their respective chromosomal numbers, scaffolding is performed. The programs utilize a reference genome, to accurately reorder highly repetitive and lower coverage regions (Mikhail, 2019; Haghshenas, 2020).

The whole genome assembly process guarantees that a large, contiguous sequence may be constructed without ambiguity. After a genome has been assembled, genomic annotation can produce data that can be utilized to conduct various kinds of research on the model organism. The whole DNA sequence of the organism is then mapped in areas relevant to the research aims. The genome sequence is named and a function is assigned during this process.

In the earlier days of bioinformatics, scientists performed computational annotation of long protein-coding genes on singular genomes per species. Recent developments enabled the scientific community to perform single nucleotides on many nucleotides at once. These developments increased the accuracy of the resolution of genome annotations, genotypes and phenotypes alike (Josep, 2019).

Common softwares and their functions

As technologies such as NGS and ONT develop rapidly, the subset of softwares that aids such scientific analysis continues to grow each year. Out of the dozens of tools available at hand, those notable and suited for the pipelines we built are discussed below.

1. Guppy

Guppy is a multipurpose base caller that works on CPUs but can also leverage GPUs for faster basecalling. Basecallers such as Guppy processes the raw signal recorded by the nanopore gathered by a sequencer and translates it into a DNA sequence. Factors such as the electrical resistance of a pore affect the base caller quality and usability due to the number of possible states growing higher. Modern base callers use a neural network model to further identify the signals with higher accuracy. The amount of training data set present in recently released basecallers also affects the quality of the basecaller (Wick et al., 2019).

2. LongQC

LongQC is a tool that identifies problems in third-generation sequencing (TGS) data using statistical methods. It works with all major TGS platforms to date, such as PacBio and ONT is designed to perform the quality assessment. Due to the indisputable trend of sequencing throughput growing dramatically, balancing speed and completeness has become a legitimate challenge even in the QC stage (Yoshinori, 2020).

3. NanoPlot

NanoPlot generates useful QC plots that show several features of sequencing data. It takes data in compressed fastq or fasta format, bam and (compressed) Guppy summary files, or multiple files of the same kind. It generates read length histograms, read length violin plots, cumulative yield plots, and bivariate graphs that compare the link between read lengths, quality scores, reference identity, and read mapping quality (Wouter, 2018).

4. Porechop

Porechop finds matching adapter sets by aligning subsets to all known adapters with an alignment scoring scheme. Afterward, it trims adapters from read ends, split reads with internal adapters, and discards reads with internal adapters (Ryan, 2018).

5. Filtlong

Filtlong hashes all the reference genome's 16-mers. Afterward, each of the input reads length and quality information assessed. After gathering all reads eligible for output, it gives each read a final score according to length, mean quality score, and window quality while setting an appropriate threshold and lists those that fail. Outputs all reads which didn't fail the standard output stream command in its input order (Ryan, 2018).

6. Flye, haslr

haslr

Haslr is a short- and long-read based assembler capable of efficiently assembling long sequencing reads into a genome. Long reads generated by TGS (such as PacBio or Oxford Nanopore) are combined with NGS reads (such as Illumina) from the same sample in this hybrid method. On a single computational node, haslr is capable of assembling huge genomes. Haslr uses a novel data structure coined "backbone graph" which models connections between short-read contigs onto long-read contigs. Experimental data suggests that this method works better for unique contigs.

To prevent repeating genomic regions creating difficulties during assembly, the implementation uses distinct short-read contigs. The backbone network is known to branch due to repetitive short-read contigs during this stage. It has also been shown to produce assemblies with a lower genome fraction than previous methods. This is because of the lack of unique short-read contigs in large regions of the genome. One of the primary drawbacks of this assembly method is that it takes a long time to locate all overlaps between input readings (Haghshenas, 2020).

Flye

Flye is a long-read based assembly algorithm that generates temporary paths in repeat graphs and constructs a more-accurate repeat graph from them. These structures called disjointigs are generated by Flye, unlike any other assembler, represent concatenations of genomic segment disjoints and are error-prone. Flye concatenates complex areas into one string. With this, it creates assembly graphs and uses reads to untangle this graph and resolves bridged repeats. Subsequently, it uses repeat graphs to resolve repeats which are not bridged by any reads using differences between repeat copies. The output will be highly-accurate contigs forming a path. On a study of downsampled (30× coverage) yeast genome, Flye had the highest sequence identity

(~99.95%). In another benchmarking experiment, Flye fully assembled 14 out of 16 chromosomes on the original ONT (120× coverage) data (Mikhail, 2019).

Difference between haslr and Flye

In one study, Flye was found to create more precise and contiguous assemblies while also providing useful information for assembly completion. Flye also reconstructs the mosaic structure of segmental duplications which is a notably difficult problem even for finished genomes (Mikhail, 2019).

7. Racon-Medaka

Racon-Medaka is a long-read based polishing tool combination. It has an error-correction stage that may be skipped, and Racon, a single instruction multiple data (SIMD)-accelerated, partial-order alignment-based, stand-alone consensus module, can construct high-quality consensus sequences quickly. Racon, when used in conjunction with a fast assembler, allows for the speedy generation of genome sequences with high precision without the need for an error-correction phase. It may be used as a rapid and accurate read rectification tool as well as a stand-alone, platform-independent consensus module for long and erroneous readings. It recovers the multi-alignment of the reads by aligning them to the genome assembly and then re-doing the consensus calculation using the original or additional read data (Vaser, 2017).

Medaka turns nanopore sequence data into variant calls with the help of a neural network. In a benchmarking experiment conducted alongside similar technologies on a 90 coverage dataset, medaka outperforms most, and when combined with Racon, its quality is slightly increased. Its performance is complemented by Flye (Oxford Nanopore Technologies Ltd., 2018).

6. POLCA

POLCA is a short-read based polishing software, implemented as a bash script program that takes as input a file of Illumina reads and the target assembly to be polished. The outputs are the polished assembly and a VCF (variant call format) file containing the variants used for polishing. The basic outline of the script is to align the Illumina reads to the genome and then call short variants from the alignments. A variant call is treated as a putative error in the consensus if the count of the alternative allele observations is greater than 1 and at least twice the count of the reference allele. Each error is fixed by replacing the error variant with the highest scoring alternative allele suggested by the Illumina reads. The variants can be substitutions or insertions/deletions of one or more bases (Zimin, 2020).

7. Ragout

Ragout is a reference based scaffolding software. It uses multiple references along with the evolutionary relationship among these references to determine the correct order of the contigs. Ragout takes an assembly as input and produces high-quality scaffolds in the form of ordered lists of contigs (Mikhail, 2014).

8. LRSDAY

LRSDAY is a multi-purpose pipeline for assembling and annotating yeast genomes from raw long- and short-read sequencing data. Long-read-based de novo genome assembly, short- and long-read-based assembly polishing, reference-guided assembly scaffolding, and extensive genomic feature annotations are all handled by LRSDAY's task-specific modules. Depending on the demands of the users, these activities can be performed separately, selectively, or in coordination (Yue, 2017).

9. GenomeQC, BUSCO

GenomeQC is a toolkit that explores genomic diversity and gene expression difficulties in a genome and produces metrics. Supported by R and Python, it utilizes a combined approach as a multi-purpose tool for the following uses: comparing reference genomes, analyzes genome assemblies and genome annotations (Manchanda, 2020).

BUSCO is a comprehensive toolkit which gives an estimate of completeness of genome assemblies. The software uses its own dataset and is compatible or non-extensive datasets, uses a prodigal gene predictor and supports batch analyses (Mosè, 2021).

2. Aims & Objectives

Saccharomyces cerevisiae strains have a vital role in both the food and alcohol industry, but may also colonize and even infect us as rare opportunistic pathogens. Thus they are in our everyday life. Their genomics is exceptionally unique and complex, their evolution is strongly connected to their domestication. Interestingly, many domesticated clades include known human isolates. Understanding their genomic and evolutionary events is essential to reaching their full potential. Through whole genome assembly and annotation based on third generation sequencing data, we are able to further understand their genomes, and carry out more precise genomic analyzes. With many tools and toolkits in bioinformatics, many scientists struggle to decide between many tools for their experiments. This experiment's aim was to compare the performance quality and time demand of two pipelines with two popular assembler tools (Flye and haslr) using long- and short-read data. Whole genome sequencing, assembly and annotation of the wine-yeast related human isolate “16406” were done to provide more elaborate results.

We expected Flye to excel in long-reads and take more time than haslr due to its polishing mechanism, which the latter does not possess. The reconstruction was expected to be done on both pipelines with minimal error and a straightforward graph resembling y=x.

3. Materials and Methods

3.1. Strain maintenance

The yeast strain used in this study was provided by the Department of Molecular Biotechnology and Microbiology. We analyzed the wine-yeast related isolate strain "16406", recovered from the trachea of a patient with Aspergillosis, and the reference genome in this experiment refers to the S288c *Saccharomyces cerevisiae* strain (Retrieved from: ***https://www.ncbi.nlm.nih.gov/assembly/GCA_002057635.1/***). In YPD broth, we revived the "16406" strain from a frozen (–70°C) culture collection, grown at 30°C overnight. A negative control was applied to test if the yeast was successfully revived.

3.2. Genomic analysis

3.2.1. DNA extraction and purification

The Zymo Research Quick-DNA™ HMW MagBead Kit was used at this step with its protocol with minor adjustments. The DNA extraction was carried out as per protocol, with the exception of 10 µl of R-Zymolease being added instead of 25 µl lysozyme. The latter part, DNA purification, was performed in accordance with the manual and the genetic material was stored at –20°C upon extraction.

3.2.2. Library preparation and sequencing

For library preparation, we used the Ligation sequencing kit (SQK-LSK109) with the 1–12 barcoding expansion kit (EXP-NBD104) as per protocol. The sequencing was carried out on a FLO-MIN106 flow cell, in a Mk1b MinION device with default settings. Basecalling was carried out post-sequencing with Guppy Basecalling Software v5.0.11 (GPU mode, high accuracy model, default options). Short-read data was retrieved in a previous study via Illumina sequencing (provided by the Department, unpublished data).

3.2.3. *de novo* Genome assembly and annotation

This experiment was run on a computer equipped with 128 GB RAM, AMD Ryzen 9 5950x - 3400 MHz, 16 cores CPU, Nvidia GeForce RTX 3090 - 24 GB memory GPU with the softwares detailed in Table 1.

Table 1: *Softwares used for the analysis in detail*

	Name	Application	Link	Reference
1	Guppy	Basecalling	https://nanoporetech.com/	(Wick, et al. , 2019)
2	LongQC, NanoPlot	Sequencing quality check	https://github.com/yfukasawa/LongQC https://pypi.org/project/NanoPlot/	(Yoshinori, 2020 ;Wouter, 2018)
3	Porechop	Adapter removal	https://github.com/rrwick/Porechop	(Ryan, 2018)
4	Filtlong	Quality control, Downsampling	https://github.com/rrwick/Filtlong	(Ryan, 2021)
5	Flye, haslr	Assembly	https://github.com/fenderglass/Flye/ https://github.com/vpc-ccg/haslr	(Mikhail, 2019; Haghshenas, 2020)
6	Racon-medaka	Long-read based polishing	https://nanoporetech.github.io/medaka/benchmarks.html	(Vaser, 2017; Oxford Nanopore Technologies Ltd., 2018)
7	POLCA	Short-read based polishing	https://github.com/alekseyzimin/masurca	(Zimin, 2020)
8	Ragout	Region-based scaffolding	https://github.com/fenderglass/Ragout/releases	(Mikhail, 2014)
9	LRSDAY	-mitochondrial genome assembly -Nuclear genome assembly finalization -Gene, TE, X&Y elements annotation	https://github.com/yjx1217/LRSDAY	(Yue, 2017)
10	GenomeQC BUSCO	Assembly quality check	https://genomeqc.maizegdb.org/ https://busco.ezlab.org/	(Manchanda, 2020; Mosè, 2021)

The pipeline (*Figure 1*) starts by base calling via Guppy. Afterwards, sequencing quality checks are done by LongQC and Nanoplot. Adapters are removed via Porechop and quality control and downsampling is done via Filtlong at 60× coverage. Assembly is now done by both Flye and haslr at this step. For Flye, a long-read based polishing is done by Racon-Medaka before both pipelines utilize short-read based polishing by POLCA. The previous step is then followed by reference-based scaffolding; Mitochondrial and nuclear genome assembly finalization, gene, TEs, core X and Y prime elements annotation were carried out with LRSDAY. *(Table 1)* Quality check of the assembly was done with BUSCO and Genome QC.

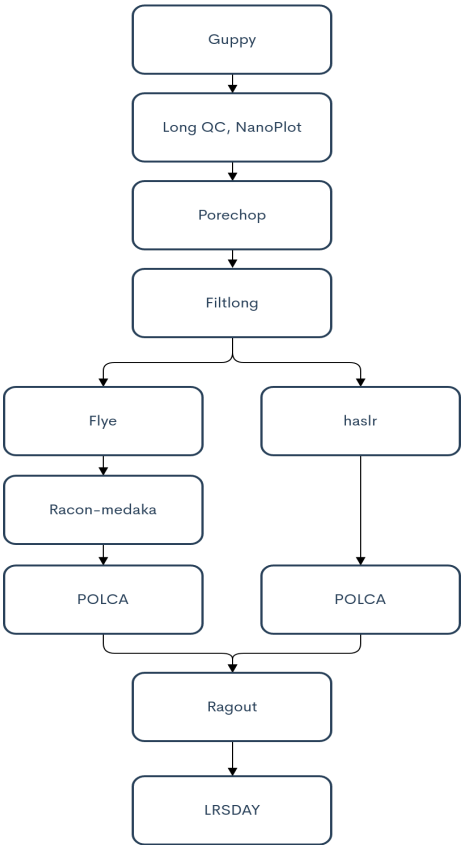


Figure 1: Pipelines optimized for short- and long-reads with its corresponding softwares (from top to bottom)

4. Results

4.1 Sequencing quality statistics

Sequencing quality statistics by NanoPlot and LongQC show that the sequencing experiment was carried out properly and the data's quality is satisfactory. Detailed statistics are in *Table 2:* our longest read is 122,796 base pairs, and all of our reads are above the Q7 quality cutoff, N50 is 19,433. Read length distribution is as expected, GC content resembles that of a whole genome. *(Figure 2*)

***Table 2**: Sequencing statistics by NanoPlot*

Mean read length	12,007
Mean read quality	13
Median read length	8,336
Median read quality	13
Number of reads	53.366
Read length N50	19,443
Total bases	640,791,007
Number and percentage of reads above Q7 cutoff	53,366 (100%)
Longest read	122,796

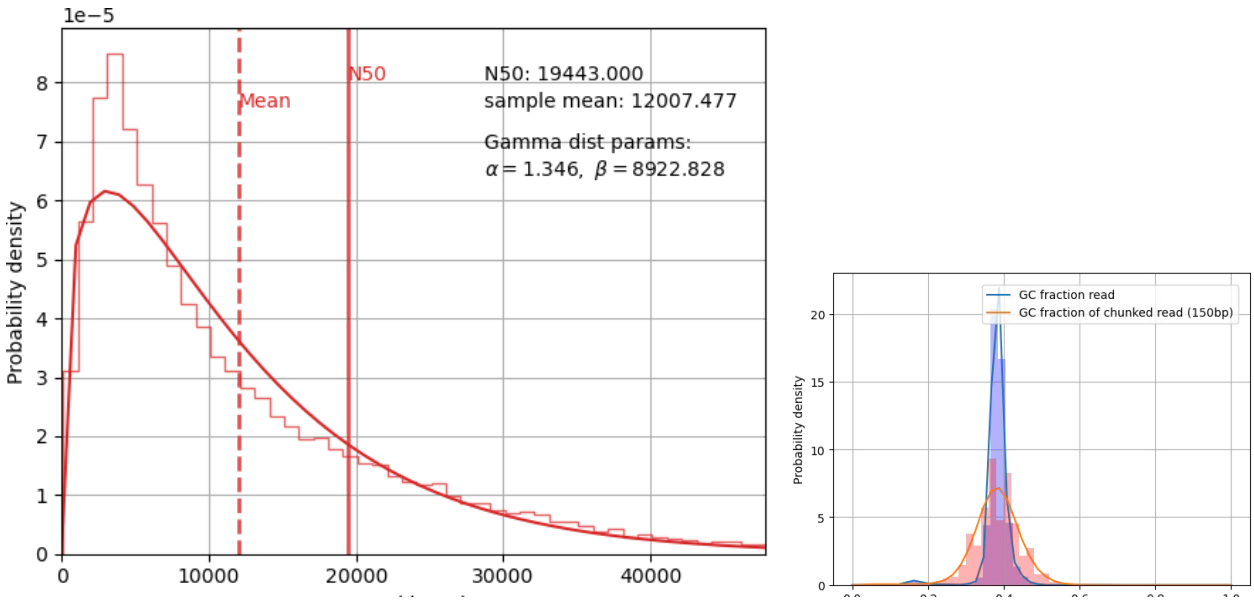


***Figure 2.:** Read length plot and GC content by LongQC.*

4.2. Assembly and annotation pipeline comparison

The graphs and chart below suggest that there is strong evidence of both pipelines being appropriate for different methods and the specific aim of the research. They vary in time required for analysis, amount of data generated and the compatibility of different types of data. We were able to generate a plot graph of the mitochondrial DNA with Flye, but not with haslr.

4.2.1 Nuclear genome assemblies

The Flye pipeline *had a longer run-time*, compared to that of haslr. Haslr produced a cleaner assembly in a shorter time (*Figure 3*). No manual curation was needed in this instance. Its precision was obstructed (*Figure 4*) in comparison with the haslr pipeline due to 3 chromosomes conjoining together during the nuclear genome assembly. Manual curation was required during the final supervised assembly as shown below.

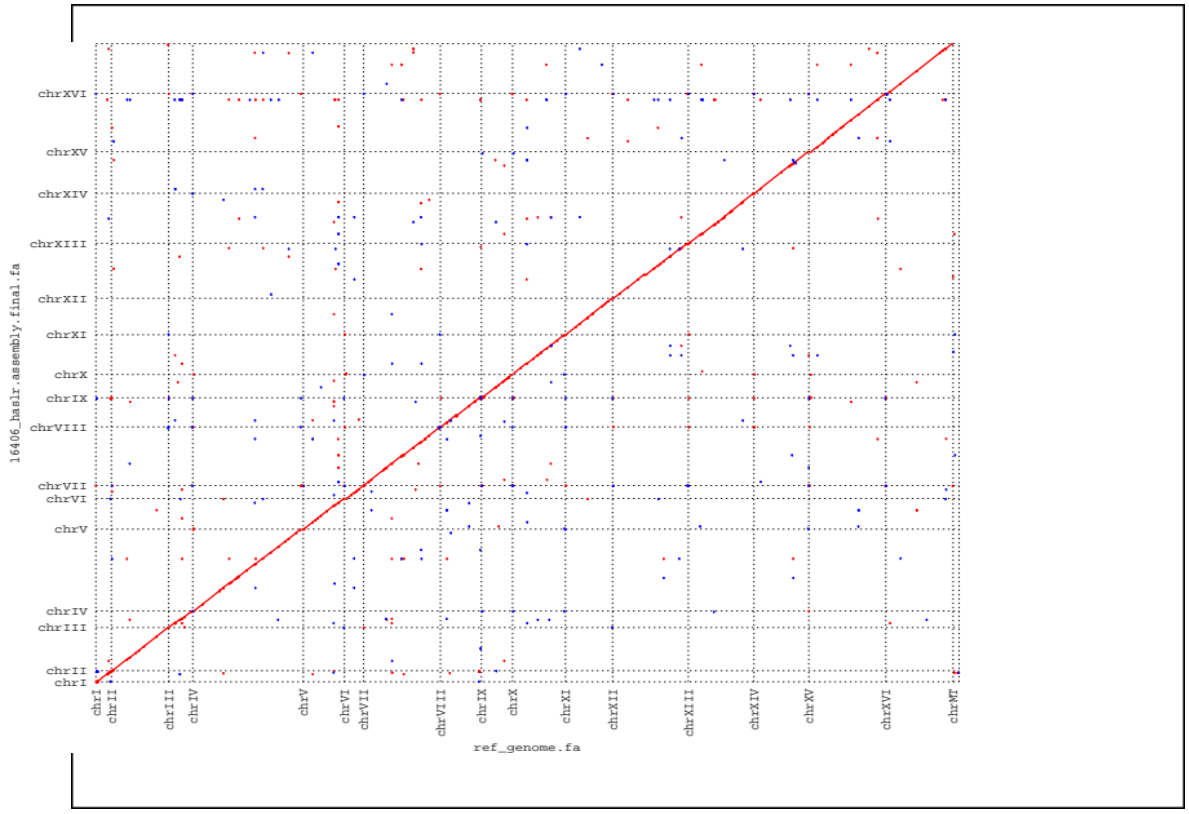
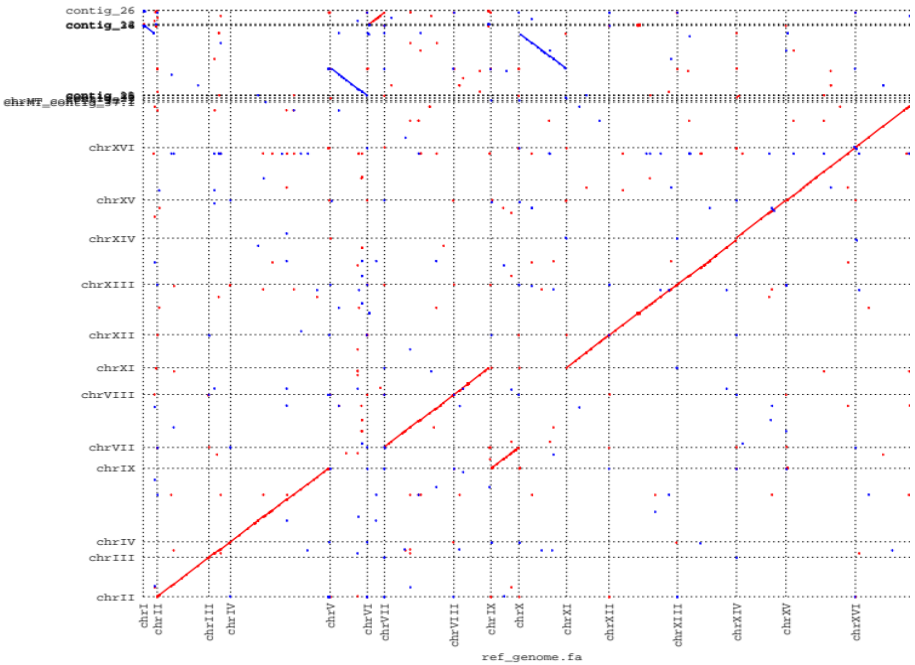


Figure 3: *The haslr generated assembly: The results show an almost-perfect y=x line.*



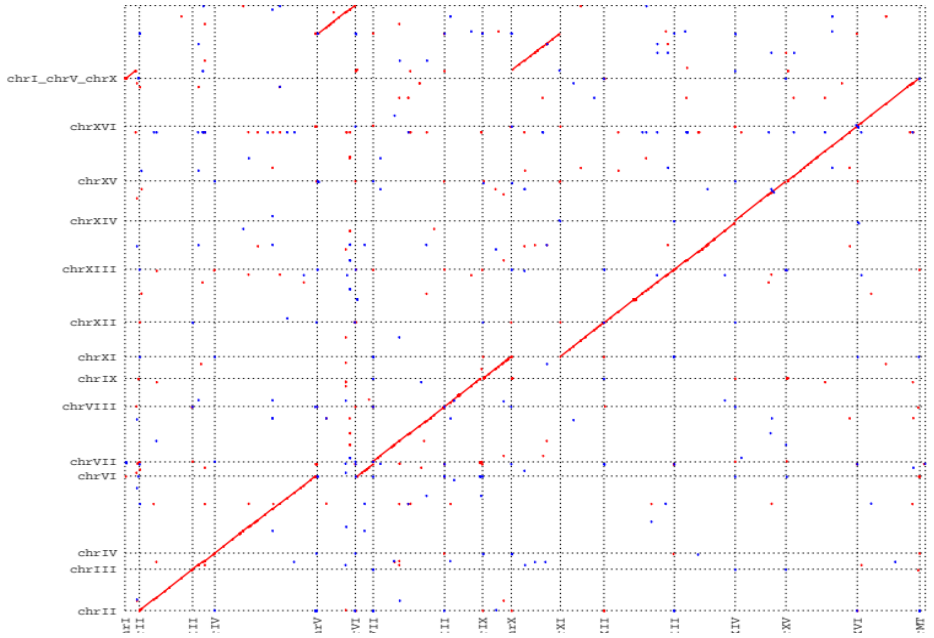


Figure 4: *Inter-chromosomal arrangement dot-plots. Two strains are compared: The S288c S. cerevisiae (X-axis); and the MinION assembly of strain "16406" (Y-axis). Homologous sequence is shown in red (forward-facing) or blue (reverse). a) Originally generated, unedited dot-plot with displaced and reversed contigs. b) Manually curated dot-plot. Rearranging chromosome VI and IX resulted in a more continuous genome. Reversed chromosomes (although conjoined), now are in the proper direction.*

After the assembly, quality check via BUSCO and GenomeQC was performed (***Table 3 and 4***).

Table 3: *Quality statistics of the nuclear DNA assembly shown with sequence lengths, counts and N50,L50,N90,L90 values as well as the makeup of the bases.*

	Statistics	Flye	Haslr
1)	**total sequence count**	14	16
2)	**total sequence length**	11730500	11443815
3)	**min sequence length**	275284	199582
4)	**max sequence length**	1493590	1467280
5)	**mean sequence length**	837892.86	715238.44
6)	**median sequence length**	863977	727679.5
7)	**N50**	1024421	891196
8)	**L50**	5	6
9)	**N90**	542884	420046
10)	**L90**	11	13
11)	**A%**	30.89	30.91
12)	**T%**	30.82	30.85
13)	**G%**	19.14	19.12
14)	**C%**	19.14	19.12
15)	**AT%**	61.71	61.76
16)	**GC%**	38.29	38.24
17)	**N%**	0	0

Table 4: *BUSCO quality check statistics generated after the nuclear genome assembly*

	Statistics	haslr	Flye
1)	**Complete BUSCOs (C)**	2128	2126
2)	**Complete and single-copy BUSCOs (S)**	2082	2081
3)	**Complete and duplicated BUSCOs (D)**	46	45
4)	**Fragmented BUSCOs (F)**	3	3
5)	**Missing BUSCOs (M)**	6	8
6)	**Total BUSCO groups searched**	2137	2137

4.2.2. Mitochondrial Genome Assembly

During the Flye assembly, the mitochondrial genome was partially assembled. (***Figure 5***) Haslr failed to assemble the mitochondrial genome.

After the assembly, we ran a quality check via GenomeQC on the mtDNA data. The AT% and GC% resemble the mitochondrial features. (***Table 5***)

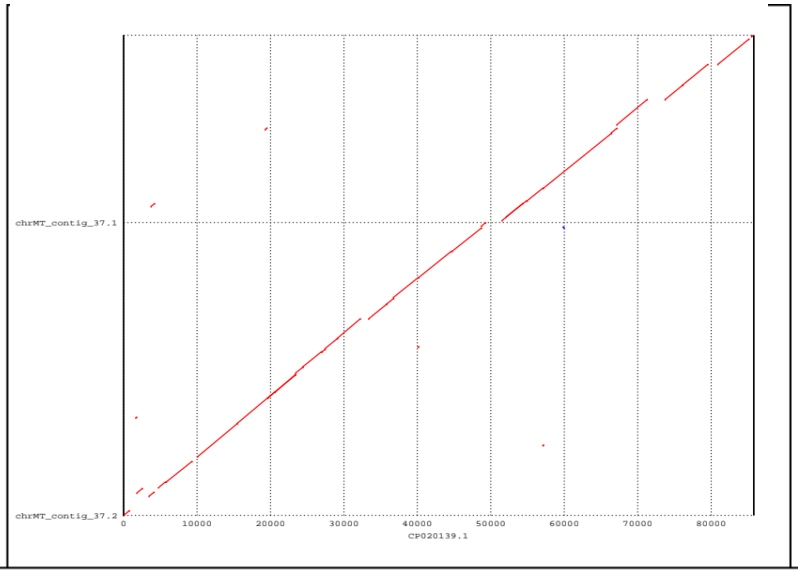


Figure 5: *The Mitochondrial assembly: Here we see minimal amounts of missing/misplaced contigs with nearly perfect alignment.*

Table 5: *Quality statistics of the mtDNA assembly shown with sequence lengths, counts and N50,L50,N90,L90 values as well as the makeup of the bases.*

	Statistics	**Flye - mtDNA**
1)	**total sequence count**	2
2)	**total sequence length**	79855
3)	**min sequence length**	31160
4)	**max sequence length**	48695
5)	**mean sequence length**	39927.5
6)	**median sequence length**	39927.5
7)	**N50**	48695
8)	**L50**	1
9)	**N90**	31160
10)	**L90**	2
11)	**A%**	42.29
12)	**T%**	41.54
13)	**G%**	8.57
14)	**C%**	7.6
15)	**AT%**	83.83
16)	**GC%**	16.17
17)	**N%**	0

4.2.3. Annotation

For the nuclear genome, both aforementioned pipelines were used. Only Flye was able to assemble the mitochondrial genome. We employed LRSDAY's annotation tools to extract data regarding the telomeric regions and TEs. We divided and compared the results by Gene count, CDSs and TE/LTR, X and Y elements. (*Appendix Table 1*)

In a gene retrieval command out of 6'500 genes, Flye was only able to recover 8. Possible conjoint chromosomes were observed on chr. I, chr. V, chr. X and chr XI's value was abnormally

higher than the reference genome. Haslr's algorithm contained its satisfactory results throughout the annotation except for core X and prime Y values.

Core X and Y prime components in haslr, on the other hand, were either missing or wrongly read. We retrieved only one value out of the 13 chromosome values, and for the core X element, only 4 out of 16 values were retrieved. In general, core X elements should be 2/chromosome, and Y' elements are expected to be identical to that of the reference genome.

For Flye, we received 14 out of 16 values with one over-representation on chr. II for the core X element whereas this was preceded by 7 out of 13 Y' elements. These shortcomings signal the pitfalls of these tools which may have to do with long- and short-read data and their characteristics.

Gene retrieval and CDS values for Flye were inadequate when based on the whole genome. *(Figure 6)* Haslr provided more outputs than Flye during TE/LTR analysis, however, the situation is reversed when it comes to core X components. *(Figure 7)* We found a significant increase in Flye findings when compared to haslr's, and unusually, haslr's Y' element output was null.

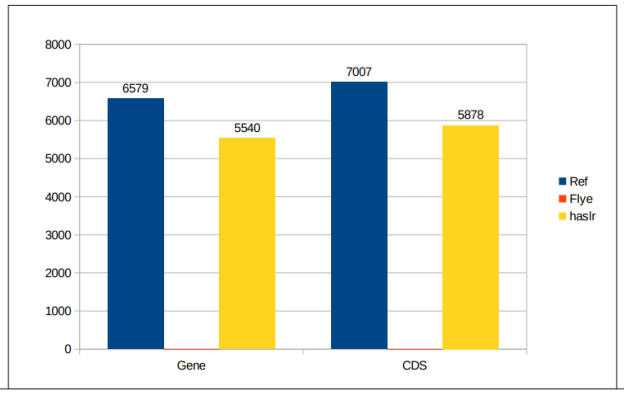


Figure 6: *Gene (Column I) and CDS (Column II) values' outputs in comparison (Individual chromosome values)*

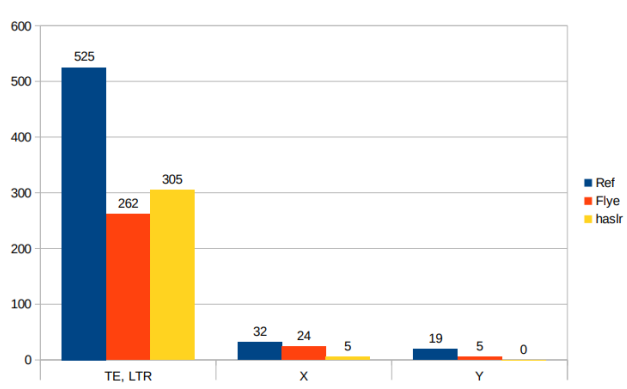


Figure 7: *TE/LTR (Column I) and core X (Column II) & Y prime (Column III) elements' output in comparison in the whole genome*

5. Discussion

Background

The process of assembling an organism's genome sequence and organizing it into chromosomes is known as a genome assembly. To establish the sequence, the gene sequence was annotated in addition. During analysis, the genome of this strain, the quality of the sample, the sequencing equipment, and the software used to put it together are all elements to consider while taking the results into account.

Baker's yeasts' haploid genomes consist of 16 individual chromosomes (and in addition, a mitochondrial genome). Its reproduction ranges from asexual to sexual, existing in several different states depending on its stage of life. This may affect repeats and more complex regions the assembly tools have difficulty tackling. The heritage of the wine yeast in particular is highly inbred with little to no variation between commercial strains and resembles the S288c more than other variants. Given these qualities, we expected the results to be somewhat highly similar to that of the reference genome.

Haslr vs Flye results

During assembly, haslr required the least amount of time to process while maintaining the highest quality, whereas Flye required a longer run-time. Flye was able to create a nearly complete mitochondrial genome dot-plot despite just processing 14 scaffolds. (Table 2) The mitochondrial genome was anticipated to have 85'000 base pairs, and we received 79'000 base pairs during assembly. Due to the Nanopore library and sequencing apparatus' presumed difficulties with mitochondrial DNA (mtDNA), the assembly usually fails to prepare the mitochondrial genome genomic data as such. The quality of the dot-plots is contrasting. Albeit the difficulty with the mtDNA, haslr produced a clearer, comprehensible graph. Flye was able to produce both the nuclear and the mitochondrial assemblies which needed manual correction in its later steps.

The BUSCO quality assessment indicates that the genome we acquired after assembly is of high quality; while an average of 80 percent is regarded as satisfactory, we received an outstanding 99.6%. This indicates that the yeast strain received a high-quality input for the majority of its 12.5 million megabases. Unreleased data from previous experiments suggest a

the median threshold score of 90 percent or above is deemed of quality. For an in-depth investigation, we looked at the correlation between the various outputs for further investigation with the aid of core X and Y prime elements.

Haslr & Flye's performance in annotation

Several oddities were discovered during annotation. Flye's shortcomings with this program appear when it comes to gene retrieval: in an array with 6'500 genes predicted, Flye only being able to recover 8. Haslr's gene retrieval hits were above average in this aspect.

In chapter 4, we discussed the possibility of 3 chromosomes conjoining together which might be one of the reasons Flye's output was of subpar quality. Haslr's 'backbone' algorithm clearly outshines here – where Flye prevails with mtDNA (Figure 6), haslr prevails with a steadfast prowess in nuclear genome assembly.

For a deeper look, we investigated its annotative results. Regarding TE/LTR values, Flye failed to show up data for chrI, chrV, and chrX, meanwhile chrXI had an over-represented value – indicating possibly misassigned chromosome(s). Haslr's outputs were never fully represented but there were no missing values.

These varying results signal a stark difference between the two assemblers: Flye may be compatible with longer reads where complex regions are low, while in turn, haslr may be the choice of tool where complex region-rich datasets are rich in short-read data.

Possible difficulties with the mitochondrial genome

To troubleshoot why only one pipeline produced mtDNA, benchmark results and experimental data from similar studies were employed. According to similar literature, a crucial step in extracting and successfully processing mtDNA could be tied to its base-calling quality. During base-calling, misreadings on modified nucleotides and homopolymer sequences can result in an error of ~10% that may affect the later stages of analysis. (Magi, A., 2016)

Another study suggests that pipeline design, specifically polishing times may affect the output somewhat. In our pipeline, both long- and short-reads are put through the polishing stage one and three times respectively, where Flye has its polishing algorithm as well. In a study where the mtDNA yield made up ~0.53% of total sequenced bases, a pipeline containing a minimum of the polishing step was repeated 5 times for clearer data. ~93.3% of mtDNA was retrieved during Flye's assembly, whereas in haslr it is not present. But this experimental data has a few

monumental differences from our study: its subject DNA sample is a vertebrate – which ONT is more-likely to have problems with, its coverage was set to ~77.5× -84× on average whereas this study performed with 60× coverage. The selection of appropriate DNA extraction methods due to these differences may have affected DNA yield and integrity, the two pillars of the result, could explain the differences in the two obtained datasets. (Mayjonade, B., 2016) This also may be a factor to consider.

Newer experimental data suggests working with two datasets, one obtained from a third generation ONT technology, another from the conservative second-generation short-read technology could enhance the quality of the data. This data was observed in the DNA samples of nematodes. (Chandler, J., 2017, Torres, L., 2018)

The main pitfall of working with one type of data is that it has problems dealing with truncated ORFs of coding genes due to premature stop codons. These mishaps happen due to modified nucleotides in polynucleotide-rich areas. This affects the base-calling step and results in misassigned bases. (Stoiber, M. H., et al. 2017) Haslr uses an approach where it draws both types of data to assemble the sequence, which may explain its success with the nuclear genome.

A quality loss resulting from a choice in polishing software is acknowledged. Racon, despite its upsides, cannot produce the same quality of reads as Nanopolish, another software that is both time-consuming and more expensive than Racon. An experimental study also suggests newer approaches such as Minimap, a tool that combines the methods of both Racon and Nanopolish to produce cost-effective results. (Senol Cali, Damla et al., 2019) Albeit making up a smaller part of the analysis, the conscious choice to use Racon may have affected the data.

6. Acknowledgements

I would like to thank the joint supervision by Lovászné Rácz Hanna Viktória, Prof. Valter Péter Pfliegler, and the Department of Molecular Biotechnology and Microbiology as well as the Talent Management Programme of the University of Debrecen (TALENTUD-DETEP).

The work was supported by the National Research, Development and Innovation Office [NKFIH FK 138910 toV.P.P.]. The supervisory work of Valter Péter Pfliegler was supported by the ÚNKP-21-5-DE-467 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

7. References

Alex N. Salazar, Arthur R. Gorter de Vries, Marcel van den Broek, Melanie Wijsman, Pilar de la Torre Cortés, Anja Brickwedde, Nick Brouwers, Jean-Marc G. Daran, Thomas Abeel. (2017) Nanopore sequencing enables near-complete de novo assembly of *Saccharomyces cerevisiae* reference strain. CEN.PK113-7D, FEMS *Yeast Research, Volume 17, Issue 7*. https://doi.org/10.1093/femsyr/fox074.

Alioto, T. S. et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. 6, 10001

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. Am. J. Hum. Genet, 73, 823–834.

Bakalinsky AT, Snow R (1990) The chromosomal constitution of wine strains of *Saccharomyces cerevisiae*. Yeast, 6, 367–382.

Barnett, J. (1992) The taxonomy of the genus Saccharomyces Meyen ex Rees: a short review for non-taxonomists. Yeast, 8, 1–23.

Benítez, T., & Codón, A. C. (2002). Genetic diversity of yeasts in wine production. Applied Mycology and Biotechnology, 2(C), 19–44. https://doi.org/10.1016/S1874-5334(02)80005-3)

Bergström A., Simpson J.T. , Salinas F. et al., (2014) A high-definition view of functional genetic variation from natural yeast genomes . Mol. Biol. Evol. ;31 :872 –88 .

Betsy Foxman, (2012). Molecular Tools and Infectious Disease Epidemiology. Academic Press . Hardcover ISBN: 9780123741332

Borneman A. R., Desany B. A., Riches D., Affourtit J. P., Forgan A. H., Pretorius I. S., Egholm M., Chambers P. J. (2011) Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of *Saccharomyces cerevisiae*. PLoS Genet.;7:e1001287. doi:10.1371/journal.pgen.1001287

Brown C. A., Murray A. W., Verstrepen K. J., (2010) Rapid expansion and functional divergence of subtelomeric gene families in yeasts . Curr. Biol. ;20 :895 –903 .

Carreto L., Eiriz M.F., Gomes A.C., Pereira P.M., Schuller D., Santos M.A. (2008) Comparative genomics of wild type yeast strains unveils important genome diversity. BMC Genom. ;9:1–57. doi: 10.1186/1471-2164-9-524.

Chandler, J., Camberis, M., Bouchery, T., Blaxter, M., Le Gros, G., & Eccles, D. A. (2017). Annotated mitochondrial genome with Nanopore R9 signal for Nippostrongylus brasiliensis. F1000Research, 6, 56. https://doi.org/10.12688/f1000research.10545.1

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet, 10(10), 691–703. https://doi.org/10.1038/nrg2640

Cubillos, F. A., Vasquez, C., Faugeron, S., Ganga, A., and Martinez, C. (2009) Self-fertilization is the main sexual reproduction mechanism in native wine yeast populations, FEMS Microbiol. Ecol., 67, 162-170.

Degré, R. (1993) Selection and commercial cultivation of wine yeast and bacteria, in Wine Microbiology and Biotechnology (ed. G.H. Fleet), Harwood Academic Publishers, Chur, Switzerland, pp. 421–48.

Dominguez Del Angel, Victoria et al. (2018) "Ten steps to get started in Genome Assembly and Annotation." F1000Research vol. 7 ELIXIR-148. doi:10.12688/f1000research.13598.1

Dunn B., Levine R.P., Sherlock G. (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. BMC Genom. ;6:1–21. doi: 10.1186/1471-2164-6-53.

Ebbert, M. T. W. et al. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 20, 97. https://doi.org/10.1186/s13059-019-1707-2

Eldarov M. A., Kishkovskaia S. A., Tanaschuk T. N., Mardanov A. V. (2016) Genomics and Biochemistry of *Saccharomyces cerevisiae* Wine Yeast Strains. Biochemistry (Mosc). Dec;81(13):1650-1668. doi: 10.1134/S0006297916130046. PMID: 28260488.

Fay J. C., Benavides J. A. (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet. 2005 Jul;1(1):66-71. doi: 10.1371/journal.pgen.0010005. PMID: 16103919; PMCID: PMC1183524.

Fay, J. C. (2013).The molecular basis of phenotypic variation in yeast. *Curr. Opin. Genet. Dev* **23**, 672–677

Fleet G. H. (1998) The microbiology of alcoholic beverages. In: Wood B.J.B. (eds) Microbiology of Fermented Foods. Springer, Boston, MA. Retrieved from: "https://doi.org/10.1007/978-1-4613-0309-1_9"https://doi.org/10.1007/978-1-4613-0309-1_9

Franco-Sierra, N. D., Díaz-Nieto, J. F. (2020) "Rapid mitochondrial genome sequencing based on Oxford Nanopore Sequencing and a proxy for vertebrate species identification". Ecol Evol.; 10: 3544– 3560. https://doi.org/10.1002/ece3.6151

Fukasawa, Y., Ermini, L., Wang, H., Carty, K., & Cheung, M. S. (2020). LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 (Bethesda, Md.)*, *10*(4), 1193–1196. https://doi.org/10.1534/g3.119.400864

Gallone B., Steensels J., Prahl T., Soriaga L., Saels V., Herrera-Malaver B., Merlevede A., Roncoroni M., Voordeckers K., Miraglia L., et al. (2016)Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. Cell. 2016;166:1397–1410.e1316. doi: 10.1016/j.cell.2016.08.020.

Goddard, Matthew & Greig, Duncan. (2015). *Saccharomyces cerevisiae*: A nomadic yeast with no niche?. FEMS yeast research. 15. 10.1093/femsyr/fov009.

Gonçalves M., Pontes A., Almeida P., Barbosa R., Serra M., Libkind D., Hutzler M., Gonçalves P., Sampaio J.P. (2016) Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. Curr. Biol. 2016;26:2750–2761. doi: 10.1016/j.cub.2016.08.040.

Gorkovskiy A, Verstrepen KJ. (2021) The Role of Structural Variation in Adaptation and Evolution of Yeast and Other Fungi. Genes (Basel). 12(5):699. doi: 10.3390/genes12050699. PMID: 34066718; PMCID: PMC8150848.

Goto-Yamamoto N., Kitano K., Shiki K., Yoshida Y., Suzuki T., Iwata T., Yamane Y., Hara S. (1998) SSU1-R, a sulfite resistance gene of wine yeast, is an allele of SSU1 with a different upstream sequence. J. Ferment. Bioeng.;86:427–433. doi: 10.1016/S0922-338X(98)80146-3.

Gutteridge, C.S. & Priest, F.G. Methods for the rapid identification of microorganisms, in Brewing Microbiology (eds F.G. Priest & I. Campbell), Chapman & Hall, London, pp. 237–70.

Haghshenas, E., Asghari, H., Stoye, J., Chauve, C., & Hach, F. (2020) “HASLR: Fast Hybrid Assembly of Long Reads”. IScience, 23(8), 101389. https://doi.org/10.1016/J.ISCI.2020.101389

Hancks, D. C., & Kazazian, H. H. (2012). Active Human Retrotransposons: Variation and Disease. https://doi.org/10.1016/j.gde.2012.02.006

Horak, P., Fröhling, S. & Glimm, H. (2016).Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. ESMO Open 1, e000094

Hosomichi K., Jinam T.A., Mitsunaga S., Nakaoka H., Inoue I. (2013) Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics, 14, 355–355.

Howe K., Clark M.D., Torroja C.F., Torrance J., Berthelot C., Muffato M., Collins J.E., Humphray S., McLaren K., Matthews L. (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature, 496, 498–503.

Illumina (2013) “An Introduction to Next-Generation Sequencing Technology.” www.illumina.com/documents/products/Illumina_Sequencing_Introduction.pdf

Jain, M. et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338

Jiang Q, Lin L, Wang T. (2008) A new model for apoptosis research: Yeast. Prog Biochem Biophys;35:361.

Johnson, Hugh. (2019) The World Atlas of Wine. 8th ed., Mitchell Beazley,

Johnston, J.R. (1990) Brewing and distilling yeasts, in Yeast Technology (eds J.F.T. Spencer & D.M. Spencer), Springer-Verlag, Berlin, pp. 55–104.

Josep F. Abril, Sergi Castellano, (2019) Encyclopedia of Bioinformatics and Computational Biology, Elsiever. ISBN 978-0-12-811432-2.

Klein SJ, O'Neill RJ. (2018) Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res. 26(1-2):5-23. doi:10.1007/s10577-017-9569-5

Langdon, Q. K., Peris, D., Baker, E. P., Opulente, D. A., Nguyen, H.-V., Bond, U., Gonçalves, P., Sampaio, P., Libkind, D., & Hittinger, C. T. (2019). Fermentation innovation through

complex hybridization of wild and domesticated yeasts. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-019-0998-8

Legras, J. L., Merdinoglu, D., Cornuet, J.-M., and Karst, F. (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history, Mol. Ecol., 16, 2091-2102.

Legras, J., Merdinoglu, D., Cornuet, J. and Karst, F., (2021) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. [online] Wiley Online Library. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2007.03266.x#b11>

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen Ba, A. N., Jones, M., Louis, E. J. (2019). *Population genomics of domestic and wild yeasts*. https://doi.org/10.1038/nature07743

Loftus, Brendan. (2003). Genome Sequencing, Assembly and Gene Prediction in Fungi. Applied Mycology and Biotechnology. 3. 65-81. 10.1016/S1874-5334(03)80007-2.

Louis, E. J. (1995) The chromosome ends of *Saccharomyces cerevisiae*. Yeast 11: 1553–1574. https://doi.org/10.1002/yea.320111604

Magi, A., Giusti, B., & Tattini, L. (2016). Characterization of MinION nanopore data for resequencing analyses. Briefings in Bioinformatics, 18(6), bbw077. https://doi.org/10.1093/bib/bbw077

Manchanda, N., Portwood, J.L., Woodhouse, M.R. et al. (2020). GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics 21, 193 https://doi.org/10.1186/s12864-020-6568-2

Matallana, E., & Aranda, A. (2017). Biotechnological impact of stress response on wine yeast. *Letters in Applied Microbiology*, *64*(2), 103–110. https://doi.org/10.1111/LAM.12677

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., … Muños, S. (2016). Extraction of high molecular-weight genomic DNA for long-read sequencing of single molecules. BioTechniques, 61(4), 203–205. http://doi.org/10.2144/000114460

Mikhail Kolmogorov, Brian Raney, Benedict Paten, Son Pham, (2014) Ragout a reference-assisted assembly tool for bacterial genomes. Oxford Academic, doi.org/10.1093/bioinformatics/btu280

Mikhail Kolmogorov, Jeffrey Yuan, et al. (2019) "Assembly of long, error-prone reads using repeat graphs" Nature Biotechnology, www.nature.com/articles/s41587-019-0072-8.

Miller JR et al. (2010) Assembly algorithms for next-generation sequencing data. Genomics; 95(6):315-27.

Mondoux M. A., Zakian V. A. (2007) Subtelomeric elements influence but do not determine silencing levels at *Saccharomyces cerevisiae* telomeres. Genetics. 177(4):2541-2546. doi:10.1534/genetics.107.079806

Mortimer R, Romano P, Suzzi G, Polsinelli M (1994) Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* from natural fermentation of grape musts. Yeast, 10, 1543–1552.

Mosè Manni, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, Evgeny M. Zdobnov. (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes, Molecular Biology and Evolution, Volume 38, Issue 10, Pages 4647–4654

Mullikin JC, Ning Z. (2003) The phusion assembler. Genome Research. 13(1):81-90. DOI: 10.1101/gr.731003. PMID: 12529309; PMCID: PMC430959.

Nakagawa, H, Fujita, M. (2018) Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci.; 109: 513– 522. https://doi.org/10.1111/cas.13505

Nijkamp J.F. , van den Broek M. , Datema E. et al., (2012) De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology. Microb. Cell Fact, ;11 :36.

Oxford Nanopore Technologies Ltd., (2018), "Medaka 1.3.4" Medaka, Github repository, https://nanoporetech.github.io/medaka/

Parapouli M, Vasileiadis A, Afendra AS, Hatziloukas E. (2020) *Saccharomyces cerevisiae* and its industrial applications. AIMS Microbiol. ;6(1):1-31. Published 2020 Feb 11. doi:10.3934/microbiol.2020001

Peter J , De Chiara M , Friedrich A et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates . Nature . 556 :339 –44 .

Pérez-Ortın J.E., Querol A., Puig S., Barrio E. (2002) Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Res.;12:1533–1539. doi: 10.1101/gr.436602.

Pollard M. O., Gurdasani D., Mentzer A. J., Porter T., Sandhu M. S. (2018) Long reads: their purpose and place. Hum Mol Genet. 27(R2):R234-R241. doi:10.1093/hmg/ddy177

Power P, Jeffery D, Rehman MA, Chatterji A, Yankulov K. (2011) Sub-telomeric core X and Y' elements in S. cerevisiae suppress extreme variations in gene silencing. PLoS One. 6(3):e17523. doi: 10.1371/journal.pone.0017523. PMID: 21437278; PMCID: PMC3060084.

Pryde FE , Huckle TC , Louis EJ . (1995) Sequence analysis of the right end of chromosome XV in *Saccharomyces cerevisiae*: an insight into the structural and functional significance of sub‐telomeric repeat sequences . Yeast;11 :371 –82 .

Pryde, F.E., Gorham, H.C. & Louis, E.J. (1997).Chromosome ends: all the same under their caps. Curr. Opin. Genet. Dev. 7, 822–828

Querol, A. & Ramón, D. (1996) The application of molecular techniques in wine microbiology. Trends in Food Science & Technology,7, 73–7.

Rieseberg, L.H. (2001). Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16, 351–358

Ryan Wick. (2021) “GitHub - rrwick/Filtlong: quality filtering tool for long reads” github.com/rrwick/Filtlong.

Ryan Wick. (2018) “GitHub - rrwick/Porechop: adapter trimmer for Oxford Nanopore reads” github.com/rrwick/Porechop. Accessed 11 Oct. 2021.

Samuel Levy and Yu-Hui Rogers. (2010) DNA Sequencing for the Detection of Human Genome Variation | Elsevier Enhanced Reader. Retrieved October 17, 2021, from https://www.sciencedirect.com/sdfe/reader/pii/B9780123749345000039/pdf

Schacherer, J., Shapiro, J. A., Ruderfer, D. M., and Kruglyak, L. (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*, Nature, 458, 342-345.

Schmitt, M. J., and Breinig, F. (2006) Yeast viral killer toxins: lethality and self-protection, Nat. Rev. Microbiol., 4, 212-221.

Schofield, M., Rowe, S.M. & Hammond, J.R. (1995) Differentiation of brewery yeast strains by DNA fingerprinting. Journal of the Institute of Brewing, 101, 75–8.

Senol Cali, Damla et al. (2019) "Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions." Briefings in bioinformatics vol. 20,4 : 1542-1559. doi:10.1093/bib/bby017

Stoiber, M. H., Quick, J., Egan, R., Lee, J. E., Celniker, S. E., Neely, R., … Brown, J. B. (2017). De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. BioRxiv, 94672, https://doi.org/10.1101/094672

Torres, L., Welch, A. J., Zanchetta, C., Chesser, R. T., Manno, M., Donnadieu, C., … Pante, E. (2018). Evidence for a duplicated mitochondrial region in Audubon's shearwater based on MinION sequencing. Mitochondrial DNA Part A, 30(2), 256–263. https://doi.org/10.1080/24701394.2018.1484116

Slotkin R. K., Martienssen R. (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 8(4):272-85. doi: 10.1038/nrg2072.

Treangen, T. J. & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36

Treu L., Toniolo C., Nadai C., Sardu A., Giacomini A., Corich V., Campanaro S. (2014) The impact of genomic variability on gene expression in environmental *Saccharomyces cerevisiae* strains. *Environ. Microbiol.*;16:1378–1397. doi: 10.1111/1462-2920.12327.

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

Vollger, M. R. et al. (2020). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* 84, 125–140

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., & Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature communications*, *7*, 11708. https://doi.org/10.1038/ncomms11708

Walther, A.C., Hesselbart, A., & Wendland, J. (2014). Genome Sequence of Saccharomyces carlsbergensis, the World's First Pure Culture Lager Yeast. *G3: Genes|Genomes| Genetics, 4, 783 - 793.*

Wick, R.R., Judd, L.M. & Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20,** 129 https://doi.org/10.1186/s13059-019-1727-y

Wightman, P., Quain, D.E. & Meaden, P.G. (1996) Analysis of production brewing strains of yeast by DNA fingerprinting. Letters in Applied Microbiology, 22, 90–4.

Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, Christine Van Broeckhoven, (2018) NanoPack: visualizing and processing long-read sequencing data, *Bioinformatics*, Volume 34, Issue 15, Pages 2666–2669, https://doi.org/10.1093/bioinformatics/bty149

Yuasa N., Nakagawa Y., Hayakawa M., Iimura Y. (2004) Distribution of the sulfite resistance gene SSU1-R and the variation in its promoter region in wine yeasts. *J. Biosc. Bioeng.*;98:394–397. doi: 10.1016/S1389-1723(04)00303-2.

Yue, JX., Liti, G. Long-read sequencing data analysis for yeasts. *Nat Protoc* **13,** 1213–1231 (2018). https://doi.org/10.1038/nprot.2018.025

Zagulski M, Herbert CJ, Rytka J. (1998) Sequencing and functional analysis of the yeast genome. *Acta Biochim Pol.*;45(3):627-43. PMID: 9918489.

Zimin, Aleksey V, and Steven L Salzberg. (2020) "The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies." PLoS computational biology vol. 16,6 e1007981. doi:10.1371/journal.pcbi.1007981

Zimmer A., Durand C., Loira N., Durrens P., Sherman D.J., Marullo P. (2014) QTL dissection of lag phase in wine fermentation reveals a new translocation responsible for *Saccharomyces cerevisiae* adaptation to sulfite. PLoS ONE. 9:e86298. doi: 10.1371/journal.pone.00862