Comparative genomics in the era of long-reads

An application on industrial yeasts

Salazar, A.N.

DOI
10.4233/uuid:90594179-e599-4371-ac63-3fa800c53cc9

Publication date
2021

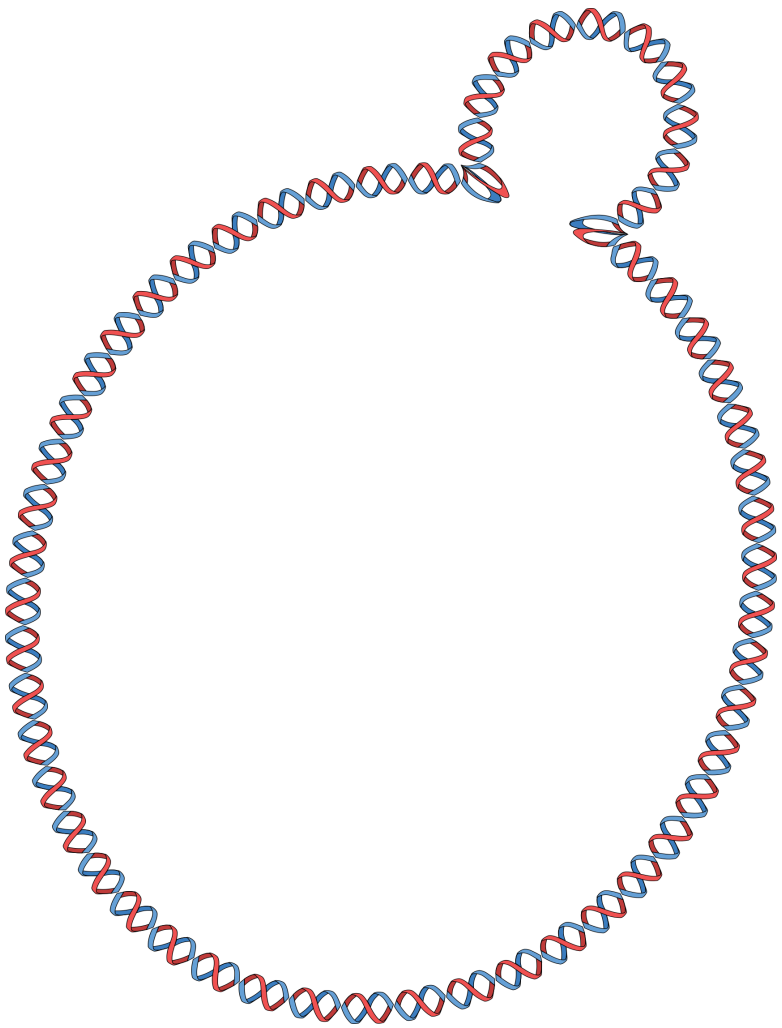
Document Version
Final published version

Citation (APA)
Salazar, A. N. (2021). *Comparative genomics in the era of long-reads: An application on industrial yeasts*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:90594179-e599-4371-ac63-3fa800c53cc9

Comparative genomics in the era of long-reads

An application on industrial yeasts

Alex N. Salazar



Comparative genomics in the era of long-reads

An application on industrial yeasts

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board of Doctorates
to be defended publicly on
Friday, 19 February 2021 at 10:00 o'clock

by

Alex N. Salazar

Bachelor of Science in Bioengineering
University of California, Santa Cruz, USA
born in San Jose, California, USA

This dissertation has been approved by the:

promotor: Prof. dr. ir. M.J.T. Reinders
copromotor: Dr. T. E. P. M. F. Abeel

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, promotor
Dr. T. E. P. M. F. Abeel,	Delft University of Technology, copromotor
Independent members:	
Prof. dr. P. A. S. Daran,	Delft University of Technology
Prof. dr. A. Schoehnhut,	Centrum Wiskunde & Informatica
Prof. dr. B. Renard,	University of Postdam
Dr. J. A. Roubos,	DSM
Prof. dr. ir. J. Fostier,	Ghent University, other member
Prof. dr. R. C. H. J. van Ham,	Delft University of Technology, reserve member

The research presented in this dissertation was funded by the BE-Basic R&D Program (http://www.be-basic.org/), which was granted a TKI-subsidy subsidy from the Dutch Ministry of Economic Affairs, Agriculture and Innovation (EL&I).

Layout by Delft University of Technology, modified by Moritz Beller

ISBN 978-90-9034-313-6

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

"The time will shortly come when the release of the complete sequence of a novel organism will no longer be a matter for excitement. The time will even come when students in biology will have difficulty in imagining that, in the obscure past, there were organisms not yet fully sequenced! How could geneticists do their work then? How could they understand what they were doing to the parts when they were missing the whole?"

—Bernard Dujon
The yeast genome project: what did we learn?
(1996)

Contents

Summary

We, humans, have an ancient microscopic companion: yeasts. These microbial organisms have helped shape our evolution, our civilizations, and our sciences. The evolutionary event that enabled yeasts to produce alcohol more than 100 million years ago was followed with adaptations throughout the animal kingdom to tolerate it. Our realisation that yeast could be used to produce bread, beer, and wine quickly enabled us to fuel the high, caloric need of many civilizations. An international dispute nearly two centuries ago about the biological nature of yeast in alcohol production, ultimately led to the founding of microbiology and the various medicinal benefits from its practice. And today, yeasts are the 'Swiss Army knives of biotechnology', as they are often engineered to produced cheaper therapeutics and alternative energy sources.

Although an ancient companion, we have only begun to truly understand yeasts and their biotechnological capabilities, largely due to a new scientific instrument: genome sequencing technology. Analogous to an 'algorithmic microscope', genome sequencing technology is enabling us to generate large amounts of data about the genetic composition and diversity of yeasts. But it comes with a challenge: these (ever-growing) datasets are complex. So how do we properly analyse them? How do we consider the complex evolutionary histories encoded in the genomes of yeasts and other microbes alike? What new biology could we learn?

The research presented in this thesis aims to provide a better understanding in the genomes of yeasts through the development and application of computational algorithms. More specifically, it focuses on two yeast species—*Saccharomyces cerevisiae* and *Saccharomyces pastorianus*—which are used in various industrial and academic institutions, either for the production of bread and alcoholic beverages, or for their genetic engineering capabilities.

I present completely new genomes for a *Saccharomyces cerevisiae* and a *Saccharomyces pastorianus* strain. They contain previously uncharacterized genes, and warrant caution in their unaccounted ability to mutate. Additionally, the genomes help test two competing theories on their evolutionary origins. I also present a visualization technique to study the evolutionary history of *Saccharomyces* genomes, and an algorithm to infer their parental origins. Addressing computational challenges when analysing microbial genomes, I also introduce a graph-based algorithm for comparing diverse genomes using a gene-centric approach.

Finally, I present a novel interactive University-level course for educating microbiologists in computational biology, helping train a new generation of scientists to navigate the world of (genomic) data.

With this thesis I have tried to stimulate your curiosity, not only in yeasts, genomics, and bioinformatics, but also in the benefits and consequences of studying the microscopic world.

Samenvatting

Wij mensen hebben een oeroude microscopisch kleine metgezel: gist. Deze microbiële organismen hebben bijgedragen aan het vormen van onze evolutie, onze beschavingen en onze wetenschappen. De evolutionaire gebeurtenis waardoor gist meer dan 100 miljoen jaar geleden alcohol kon produceren, werd gevolgd door aanpassingen door het dierenrijk om het te tolereren. Ons besef dat gist kan worden gebruikt om brood, bier en wijn te produceren, stelde ons al snel in staat de hoge, calorische behoefte van veel beschavingen te voeden. Een internationaal geschil bijna twee eeuwen geleden over de biologische aard van gist bij de productie van alcohol leidde uiteindelijk tot de oprichting van microbiologie en de verschillende medicinale voordelen van de praktijk. En tegenwoordig zijn gisten de 'Zwitserse zakmes van de biotechnologie', omdat ze vaak worden ontworpen om goedkopere therapieën en alternatieve energiebronnen te produceren.

Hoewel het een oude metgezel is, zijn we gisten en hun biotechnologische mogelijkheden grotendeels pas echt gaan begrijpen dankzij een nieuw wetenschappelijk instrument: genoomsequentie-technologie. Analoog aan een 'algoritmische microscoop' stelt de genoomsequentie-technologie ons in staat om grote hoeveelheden data te genereren over de genetische samenstelling en diversiteit van gisten. Maar dit stelt ons voor een uitdaging: deze (steeds groter wordende) datasets zijn complex. Dus hoe analyseren we ze goed? Hoe beschouwen we de complexe evolutionaire geschiedenissen die zijn gecodeerd in de genomen van zowel gisten als andere microben? Welke nieuwe biologie kunnen we leren?

Het onderzoek dat in dit proefschrift wordt gepresenteerd, heeft tot doel een beter begrip te bieden van de genomen van gisten door de ontwikkeling en toepassing van computationele algoritmen. Meer specifiek richt het zich op twee soorten gist—*Saccharomyces cerevisiae* en *Saccharomyces pastorianus*—die worden gebruikt in verschillende industriële en academische instellingen, hetzij voor de productie van brood en alcoholische dranken, hetzij vanwege de mogelijkheid tot genetische manipulatie.

Ik presenteer volledig nieuwe genomen voor een *Saccharomyces cerevisiae* en een *Saccharomyces pastorianus*-stam. Ze bevatten voorheen niet-gekarakteriseerde genen en verdienen voorzichtigheid wat betreft hun onverklaarde vermogen om te muteren. Bovendien helpen de genomen twee concurrerende theorieën over hun evolutionaire oorsprong te testen. Ik presenteer ook een visualisatietechniek om de evolutionaire geschiedenis van *Saccharomyces*-genomen te bestuderen, en een algoritme om hun ouderlijke oorsprong af te leiden. Om computationele uitdagingen aan te pakken bij het analyseren van microbiële genomen, introduceer ik ook een op grafieken gebaseerd algoritme voor het vergelijken van diverse genomen met behulp van een gencentrische benadering.

Ten slotte presenteer ik een nieuwe interactieve cursus op universitair niveau voor het opleiden van microbiologen in computationele biologie, waarmee een nieuwe generatie wetenschappers kan worden opgeleid om door de wereld van (genomische) gegevens te navigeren.

Met dit proefschrift heb ik geprobeerd je nieuwsgierigheid te prikkelen, niet alleen naar gisten, genomica en bioinformatica, maar ook naar de voordelen en gevolgen van het bestuderen van de microscopische wereld.

Preface

The topic of beer—and alcohol in general—often carries a comical connotation. However, its history and influence in human civilization is of no laughing matter. Beer can be traced back ~14,000 years ago, and has then integrated as a global cultural staple. Today, it is a billion-dollar industry, while being one of the world's most abused drugs.

But this thesis is not exactly about beer. Instead, it centers around the organisms that made beer possible, *yeast*. Specifically, *Saccharomyces* yeast. As we will shortly see, it is these organisms that brought alcohol into the animal kingdom, influenced our evolution, and propelled the field of microbiology.

Despite more than a century-worth of scientific research, we have only recently unravelled the global diversity of *Saccharomyces* yeasts, enriching our grand pursuit of using these organisms in industrial applications. These insights have been driven largely by rapid innovations in *genome sequencing technologies*, which provide deeper understanding about the genomes and evolution of *Saccharomyces* yeast. However, the data generated by these technologies, and its subsequent biological interpretations, are complex.

This thesis focuses on the development and application of *bioinformatic* algorithms that aid in our understanding of the genomes of *Saccharomyces* yeasts. Specifically, *sequence analysis* and *comparative genomics* of *Saccharomyces* genomes.

But before diving into the world of bioinformatics (and the main contents of the thesis), I wanted to expand on the influential role that yeasts have played in our modern lives. As such, **this Preface provides an overview of the history of alcohol in the animal kingdom and human society, serving as an appreciative and educational take on the historical influence of yeasts**.

So to start, let me tell you a story about an outdoor wine bar, for chimpanzees.

0.1 A brief history of beer

There is a *magical* tree found in the western coast of Africa. They are called, *Rafia* palm trees, and are at the center of an ancient ritual practiced by many of the locals. For this ritual, you must cut a hole through a Rafia palm tree (either through the main trunk, or one of the branches) and use a small container to collect the milky-like sap that starts to ooze out. The sap is usually sweet with a coconut-like taste, which you can drink right away. But instead of drinking it, leave the container open and return again in a few hours, for this is when the magic happens: the sap turns alcoholic.

Depending on your patience, the sap, now known as *palm wine*, can have an alcohol content of 3.1-6.9% [1]. If you know your Belgium-Dutch beers, the alcohol content by volume (ABV) ranges from lager-lemon version of an Amstel Radler (3.0% ABV) to a Westmalle Dubbel (7.0% ABV). In other words, the ABV in the sap ranges from a typical "weak" to a "strong" beer. But do make sure to eventually collect the container, or else unlikely group of visitors will call "dibs" on the palm wine: chimpanzees.

0.1.1 A taste for alcohol

From 1995 to 2012, a group of researchers followed a community of chimpanzees in Bossou, Guinea, whose territory overlapped with palm trees that were frequently "tapped" by locals to produce palm wine [1]. To their surprise, they found that the chimps had a natural taste for the alcoholic beverage. Throughout the 17 years, the researchers managed to characterize 20 different *drinking sessions* where sub-groups of chimpanzees would visit tapped palm trees and drink from the containers. Sometimes it was a lonely individual, such as on February 5, 2004, when a male chimpanzee drank 1.57 litres of palm wine by himself in a period of 17 minutes [1] (roughly equal to three pints at your local bar). Other times it was a party, such as on July 22, 2004, where a total of eight different chimpanzees (three males and five females) drank together for an unknown quantity [1]. And occasionally, it was the *usual suspects*, when a trio of chimpanzees routinely visited the containers together [1]. Although there were no breathalyzers around, some chimpanzees appeared to be intoxicated after their drinking sessions [1].

Although not exactly a wine bar for chimps, the pre-tapped palm wine containers ultimately functioned as one. And as comical as these observations may sound, they do raise two important questions: what is the *magic* behind palm wine? And is there a natural preference for alcohol in the animal kingdom?

We now know that it's not exactly magic that transforms sap into its alcoholic version: instead, the transformation is made possible by a group of microbial organisms known as *yeast*. Yeast (specifically from a group of organisms known as *Saccharomyces*) are fungi that are about 5-10x smaller than the width of a human hair, and can be found all over the world, especially in areas harbouring sugary foods such as fruits. In the case of sap from Rafia palm trees in Bossou, wild yeast on the trees, in the air, and/or on the containers left over from previous batches, mediate chemical reactions to convert sugar into ethanol (alcohol). Yeast thus produce the alcohol content in the palm wine, whose strength depends on the amount of sugars in the sap and the duration in which the yeast can mediate the chemical reactions.

Interestingly, this special ability to convert sugars into ethanol—termed, *alcohol fermentation*—seems to be largely unique to yeast [2, 3]. In other words, alcohol fermentation

is a unique ability that has (so far) only been found in a minor fraction of the ~10,000 microbial species characterized thus far [4]. However, there are some bacteria with fermentation capabilities. The bacteria, *Zymomonas mobilis*, can also convert sugar into ethanol [5]. In fact, this bacterium contributes (in minor quantities) to the alcohol content in palm wine, as well as tequila and the ancient Mexican drink, *pulque*, which are similarly fermented from the sugary sap of agave plants [5]. Overall, it's an appreciative realization: the thousands of beers, wines, sakes, whiskeys, bourbons, vodka, gins, and other alcoholic beverages are all largely dependent on the alcoholic fermentation capabilities of only a few microbes.

So how did they gain this special ability?

One hypothesis proposes that alcohol fermentation originally functioned as a *competitive mechanism*. Glucose (sugar), which can be derived from carbohydrate foods, is the main energy source for many animals and microbes. After breaking down glucose into a chemical called, *pyruvate*, organism can derive a large source of energy by digesting pyruvate through *aerobic respiration*, a chain of additional chemical reactions that require oxygen. However, when oxygen is not present, pyruvate can be digested through an alternative chain of chemical reactions termed, *anaerobic respiration*. Although the exact details of its evolution are still unclear, early ancestors of modern-day yeast evolved to have alcohol fermentation as an anaerobic system as late as ~125 million years ago [3, 6]. In other words, whenever oxygen was absent, yeast could derive energy by using pyruvate to produce ethanol. Other organisms such as some bacteria and animal muscle cells also possess an anaerobic system called, *lactic acid fermentation*, where pyruvate can be used to produce lactic acid. You likely already experienced lactic acid fermentation: lactic acid itself is produced by the bacteria *Lactobacillus* and *Streptococcus* are used to process milk into cheese and yogurt, while muscle soreness during/after an exercise session can be attributed to the build-up of lactic acid produced by muscle cells.

Importantly, both lactic acid and ethanol are toxic to many organisms. Lactic acid—as the name implies—is acidic, creating an ionic imbalance in the environment that can denature many crucial proteins in cells. Similarly, ethanol is also toxic, as once absorbed by a cell, it can chemically react and damage DNA and proteins. Thus, yeast and some bacteria have a competitive advantage as they can kill other organisms in their nearby surroundings, reducing competition for space and resources. But this competitive edge comes at a cost: the amount of energy that can be derived from anaerobic respiration is 19x lower than that of aerobic respiration [3]. So from an energy perspective, it's much more preferable to use aerobic respiration than anaerobic. However, yeast managed to find a way to do both.

About 125 million years ago, fruits began to evolve from plants, resulting in an abundant source of sugary nutrients to not only animals, but also to microbes [6]. It is around this time that three different lineages of yeast independently evolved the so-called, *Crabtree effect*: the ability to perform both aerobic and anaerobic respiration [3, 6]. More specifically, yeasts would normally derive energy via oxygen, but when there were high concentrations of sugars, they could switch to alcohol fermentation and release ethanol into the environment, giving them a huge competitive advantage for resource and nutrients. As such, yeasts were now able to "ferment" sugary foods like sap, nectar, and fruits with alcoholic content. But the toxicity of alcohol did not stop other organisms from in-

dulging on these newly fermented resources.

As a wise man once said, "Life…uhm…finds a way" [1]. And indeed, in nature we find organisms that have evolved systems to handle the toxic properties of ethanol, enabling them to make use of the sugary nutrients in alcoholic foods. One of the best examples are fruit bats, which frequently feed on fermented fruits. A study in 2010 found that fruit bats often fly with a blood alcohol content (BAC) of more than 0.3%, without any observable issues [7]. To put it in perspective, the legal BAC limit for automobile drivers in many European countries is 0.05%; in the USA it is 0.08%. This means that bats are flying under the influence at more than 4-6 times the capabilities of humans. Another example are Tree shrews, which constantly feed on fermented nectar from flowers. A study in 2008 found that the amount of alcohol ingested in tree shrews is equivalent to an average adult female drinking 9 glasses of wine in a period of 12 hours [8].

Yes, humans are not the only alcoholics in the animal kingdom.

But from an evolutionary perspective, it shouldn't be much of a surprise: two organisms whose diet primarily depend on fermented foods have the capability to ingest high amounts of alcohol. But primates (including humans) have diverse diets and are not dependent on alcoholic foods. So why—and how—did we develop an affinity to purposely seek out alcohol, sometimes in excess amounts?

Most explanations regarding our natural taste for alcohol remain speculative. But the logic is similar to fruit bats and tree shrews: early primates likely came across (overly-)ripped fruits that were fermented, and the ability to process ethanol allowed them to include these foods into their diets. Indeed, a study in 2015 showed that the last common ancestor of human, apes, and primates harboured a functional version of the gene, *alcohol dehydrogenase* [9], which is one of the main genes that allows us to process ethanol into a less toxic form. The researchers managed to do this by comparing the DNA sequence of alcohol dehydrogenase genes across different apes and monkeys, and attempted to trace back all the mutations that occurred throughout its evolution, until it converged into a single ancestral version. This ancestral version dated back around 50 million years ago, about the same time as the last common ancestors between humans, apes, and primates [9]. By inserting this ancestral version into a bacterium, they were able to express its protein and measured its ability to process ethanol. Functionally, it wasn't that great at processing alcohol, but it did its job [9].

Now, the *drunken monkey* hypothesis suggests that evolution favoured early hominid species that were attracted to ethanol [10]. Although still debated, it argues that alcohol may have provided survival advantages by serving as: a proxy to find fruits; acted as stimulants to our appetite increasing out caloric intake; and encouraged more social behaviour. But ~10 million years ago, the version of the alcohol dehydrogenase gene in the last common ancestor of humans and the great apes underwent a series of mutations that made it 40x more efficient at processing ethanol [9]. In other words, this ancient hominid species in which human and apes evolved from, were now able to process alcohol at much larger quantities. Consequently, this also meant that chimpanzees and gorillas were able to ingest alcohol in comparable levels to humans.

[1] Dr. Ian Malcom, *Jurassic Park*

Which brings me back to the significance of the Bossuou chimpanzees drinking palm wine: it was the first time that great apes (other than humans) were deliberately observed and quantified to drink alcohol in the wild. Of course, it would've been more impressive if the chimpanzees themselves were the ones tapping the Rafia palm trees of palm wine, so we do have to acknowledge that their drinking affinity is a direct consequence of human involvement. Furthermore, it's unclear whether the chimpanzees sought out palm wine for pure enjoyment, or if it served as a "fall-back" food due to limited resources. However, green monkeys in the island of Saint Kitts in the Caribbean Sea have been observed to constantly sip on cocktails of tourists. But whatever their reason may be, these studies do show that apes similarly have a natural affinity towards alcohol.

Which brings me to a particular type of alcohol that humans have become very fond of: *beer*.

0.2 The evolution of beer

Perhaps it was due to our early experiences with fermented fruits, and the way alcohol made us feel. Or perhaps it is indeed hard-wired in our genetics. Regardless of the reason, humans love alcohol. And evident from the trillions of litres of beer annually consumed around the world [11], humans particularly love beer. So, between ~10 million years ago and present day, where and how did beer originate?

Well, this question is knotted to an ancient riddle:

Which came first: bread, or beer?

Cereals (such wheat, grains, oats, legumes, and barley) are historically—and continue to be—a major food staple in human civilization. Importantly, cereals are the precursors for making bread and beer. In fact, both bread and beer are based on the same principle: extract sugars from the seeds of cereals to allow yeast to digest them into ethanol and CO_2.

Seeds are portable starting kits with all the necessary nutrients to germinate a plant. Within these nutrients are starches: large chains of glucose that are chemically linked together. Additionally, seeds also contain two proteins, *alpha* and *beta amylase* enzymes which can break down starches into different types of sugars [12]. Both of these enzymes become active when the seed is ready to germinate, and harbour different functionalities: alpha amylase randomly cleave starch molecules, producing a mixture of sugars such as glucose, *maltose*, and *maltotriose*; while beta amylase progressively cleave (or *nibble*) from the ends of the starch molecules, producing mostly maltose [12]. As such, bread dough and wort—a soupy mixture of water and mashed cereals used as the starting ingredients to brew beer—are largely made up of this sugary mixture, enabling yeast to produce ethanol and CO_2.

Indeed, modern-day yeast (especially *Saccharomyces cerevisiae*) have been specifically adapted for different types of breads and beers. For example, the appropriate yeast strain in bread-making depends on the dough being fermented, such as lean, sweet, and frozen dough [13]. The main fermentable sugar in lean dough is maltose, since the sugar composition primarily originates from the cereals used when making the dough [13]. Lean dough thus requires yeast to not only properly utilize maltose to produce CO_2 and make dough rise, but also avoid a *lagging phase* that some yeast experience when breaking down the

sugar, which leads to a drop of CO_2 production during the first hour of fermentation [13]. In contrast, sweet dough (as the name implies) has additional sugars, where up to 30% of sugars added is sucrose. This creates a high *osmotic* pressure for yeast (a pressure induced by the difference in the internal and external sugar concentrations) which can decrease yeast's fermentation ability [13]. As it turns out, some yeast can tolerate higher osmotic pressure, enabling them to better ferment sweet dough [13].

Similarly, many beer strains are better able utilize maltose and maltotriose, as these two sugars make up more than 50% of the sugars in wort [12]. Wine and cider yeast have been adapted to better utilize fructose during fermentation, since fructose is the main sugar in fruits (such as grapes and apples) [14]. Furthermore, wines usually start with higher concentrations of sugar, consequently leading to much more ethanol production. As such, wine yeast have higher tolerance to ethanol than beer yeast [15].

Nevertheless, bread, beer, and wine yeast are generally all the same species, and can be substituted for one-another when making either food. Sure, the end-product may not be "optimal" (such as the presence of "odd flavours" and low-quality beer/bread), but for yeast, as long as sugars are present, they can produce ethanol and CO_2.

So, when early humans first began to harvest wild cereals, did they originally do so to make bread, or beer?

The birth of agriculture is generally credited to the *Natufians*, a group of hunter-gathers that transitioned to farming more than ~14,000 years ago in the Near East (around modern-day Israel, Jordan, Palestine, and Syria) [16–19]. Archaeological evidence show that Natufians were among the first to harvest wild cereals to produce food, including bread and an ancient version of beer [16–19]. In fact, the ancestors of modern-day wheat and barley has been linked to the Near East [20, 21]. As such, some researchers believe that Natufians first harvested wheat for bread making, and after some serendipitous events, discovered that they could use the same cereal ingredients to brew beer [18]. However, researchers in 2018 came across stone mortars in a Natufian graveyard with chemical traces of ancient beer dating back more than ~13,000 years ago [20]. Specifically, they found high traces of small cereal compounds such as starch granules, phytoliths, and fibres, suggesting that Natufians used these mortars to crush cereals [20]. But the altered morphology of the cereal compounds highly resembled the morphology induced via alcohol fermentation, suggesting that Natufians were actually using these mortars to brew ancient beer [20]. Furthermore, the researchers showed that it's quite simple to make beer using the stone mortars: mix cereals with water, mash them, and let wild yeast ferment the gruel-like mixture into an ancient version of beer [20]. The simplicity for ancient beer thus raised questions on whether Natufians first invented beer, and later stumbled upon bread making.

The notion that beer predates bread is not new, as Dr. Robert Braidwood nearly 70 years ago first proposed the *beer hypothesis*: beer brewing was discovered first, and that our love for it motivated us to domesticate cereals, later leading to bread production [22]. However, it remains unclear which of the two came first, as a separate group of researchers in 2018 similarly came across traces of ancient bread making dating back around the same time as the ancient beer residues in the stone mortars in Natufian territory [18]; further complicating the bread or beer riddle.

Regardless of the order, the invention of beer is credited to the Natufians (at least the first archaeological instance of it). And since then, the evolution of beer (and alcohol beverages in general) was likely shaped by the combinations of independent discoveries of fermentation, along with movements of human populations.

What was well documented was the love for beer in ancient Sumer (around modern-day Iraq and Kuwait) and ancient Egypt, roughly 6,000 to 3,000 years ago. Sumerians loved beer, making it a central commodity in their economy [23]. One of the oldest writing-tablets ever recovered is a ~5,000 year-old Sumarian 'beer payslip' recovered in modern-day Uruk, Iraq, documenting beer rations paid to workers [24]. In Sumerian mythology, there was *Ninkasi*, the ancient goddess of beer. And to celebrate her, they had a poem called, *The Hymn of Ninkasi*, describing not only her origins from a sacred lake, but also outlining a Sumerian recipe for beer, via the combination of local cereals with honey [25]. In fact, this outline covers the three basic steps of modern-day beer-brewing: malting, mashing, and fermentation. Furthermore, archaeological text shows that there were at least 19 different types of beers that the Sumerians brewed: eight from wheat, eight from barley, and three made from mixture of the two [26].

Ancient Egyptians were also major beer drinkers, likely influenced by their Sumerian neighbours. In their mythology, human existence is, in part, credited to beer: after a regrettable decision by the Egyptian god, *Ra*, to summon the goddess warrior, *Sekhmet*, to destroy humanity, he tricks Sekhmet into drinking large quantities of beer, who drunkenly falls asleep to later wake up as the goddess, *Hathor*, who was ultimately kinder to humanity [27, 28]. This event was commemorated by the ancient Egyptians as the *Festival of Drunkenness*, where Egyptians would 're-enact; Hathor by drinking large amounts of beer (and wine) until they fell asleep [28]. In brewing practice, Egyptian and Sumerian differed: it's suggested that ancient Egyptians first baked bread in low temperatures (which in hindsight, allowed yeast cells to survive in the bread), crumbled it and added it to water vessels, where the yeast would then ferment remaining sugars [29]. The resulting pale, yellow beverage was referred to as *bouza* [29].

Around the same time, (Northern) Europeans were enjoying sweet versions of ancient beer. Potteries from Scotland, including the Isle of Arran and Rhum, have been found to contain traces of mashed cereals along with honey and meadowsweet (a type of herb), dating ~4,000-5,000 years ago [30]. In Egtved, Denmark, a wooden bucket was discovered at the graveyard of a woman dating back around ~3,000 years ago, which similarly contained traces of mashed cereals along with honey and berries [30, 31]. The chemical traces of these archaeological artefacts suggest a practice of ancient beer in these regions, which appear to be sweet and fruity, either precursors or paralleling mead (fermented honey), which was a common alcoholic beverage drank by Vikings and Germanic tribes [2].

Sadly, beer fell out of fashion in the Greek and Roman empires. Instead, wine dominated various regions in Europe during this time [30, 31]. Much of the negative views on beer originated from pseudo-scientific beliefs. In ancient Greece, wine was described as a 'hot' and 'dry' beverage; contrast to beer which was 'cold' and 'wet' [30]. A the same time, Greek physicians believed that males were naturally 'hotter' and 'dryer' than females [32]—likely influenced from Hippocrates' work of the *four humors* of the human

[2] Max Nelson has a fantastic in-depth historical take of beer in ancient Europe [30]. Many of the points in this sub-section are thus summaries of his work.

body [33]. Thus, wine was viewed as a masculine drink, contrast to beer which was viewed as feminine [30]. Furthermore, Theophrastus—the successor of the famous philosopher, Aristotle—believed that beer fermentation was due to the spoilage of cereals, as opposed to wine fermentation which was a "natural" transition from grapes [30].

Importantly, Gallic and Germanic tribes—who were constantly at war with the Romans—continued their practiced of beer brewing, despite the wine-influence of their Roman neighbours. Particularly, the Southern Gales (around modern-day France) brewed two main types of beer: *korma* (barley beer) and *cervisia* (wheat beer) [30]. Although different versions existed for both, such as those with honey, Southern Gales viewed wheat beer as superior to barley beer. Indeed, the Romans referred to these 'barbaric drinks' using the same Gallic name, though different variants of the names existed, such as *cervesa* [30]. It is also no coincidence that the yeast species commonly used for beer and break making is named, *Saccharomyces cerevisiae*, a Latin form of this Gallic word. Southern Gales, as well as Celtiberians and Lusitanians of the Iberian Peninsula in modern-day Spain and Portugal, natively brewed their own versions of barley and wheat beer, termed, *celia*, *caelia*, and *cerea*—but they ultimately integrated wine into their culture after being conquered by the Roman Empire roughly ~2,000 years ago[30]. The exception were the Northern Gales (around modern-day Belgium) who Julius Ceasar noted to be 'the bravest of their tribes', largely due to their rejection of Roman luxuries, which they believe made soldiers effeminate [30].

But the influence of beer on ancient empires was not restricted to the 'old world'.

First excavated in 1989, Cerro Baul—a 600-meter-high promontory in Southern Peru—was a political outpost by the *Wari Empire*, who reigned the region ~1,000-1,500 years ago [34, 35]. The site likely mediated political talks from their Southern rivals, the *Tiwanaku*. Interestingly, this political outpost housed one of the largest ancient breweries discovered: a 500 square-meter facility that brewed different variants of *chicha*, a South American beer made of maize and pepper berries [34, 35]. The facility had all the necessary infrastructure to brew large quantities of beer, housing specialized rooms for grinding, boiling, fermenting, and storing. Remarkably, several vessels in the fermentation room were found to hold up to a 150 L of liquid, with one possibly holding 1,000 L [34, 35]. It is estimated that the facility could produce up to 1,800 L of chicha per batch (that is about 5,455 standard bottles of beer). Ultimately, this large brewing facility reflected the political mindset of the Wari Empire, which held large (drunken) festivals to commemorate political agreements [34, 35].

After the fall of the Roman Empire, and into the Medieval Europe (about ~1,500 years ago), a series of events ultimately changed beer into the alcoholic drink that we love (and hate) today.

The first, was the European-wide adoption of beer brewing by Christian monasteries. This was largely due to (sequential) work from Gildas the Wise and St. Columban, who established formal monostatic rules in Ireland and Britain that not only advocated for clean and sanitized brewing practices, but also regulated the amount of beers monks could drink [30]. For example, Gildas would have monks stand still for three hours at night reciting

more than twenty-eight psalms if they were caught drunk [30]. St. Columban punished monks who spilled beer by having them recite 12 psalms; or for more severe spills, would have the monks go sober (no drinking anything but water) for a number of days equivalent to "the amount of alcohol spilled" [30]. It is also during this time where we start to see incidents of *beer miracles* by various monks and saints, ranging from unlimited beer, to spontaneous fermentation, and the equivalent of a beer exorcist.

It was not until the reign of King Louis the Pious of France when beer brewing was officially regulated throughout all Christian monasteries in Europe [30]. King Louis the Pious followed the footsteps of Charles the Great, who fortified beer brewing and wine making in his estates throughout France around 1,200 years ago [30]. These regulations forced Christian monasteries to reflect Louis' *modern standards*, such as the St. Gall monastery in modern-day Switzerland which was renovated to house three different brewing rooms for monks, special guests, and travelers such as pilgrims [30]. Nevertheless, this official regulation ultimately encouraged monks to experiment with brewing recipes. And it's during this time when we start to see the integration of a major modern-day beer ingredient: *hops*.

As previously discussed, ancient brewers have historically used various ingredients to flavour their beer, including honey, berries, and herbs. But none were as revolutionary as the additive ingredient of the herb called hops. Hops are "climbing plants" that can grow 10 meters high with three main species: *H. lupulus*, *H. japonicus*, and *H. yunnanensis* [36]. Although their origins are linked to East Asia, hops have naturally grown throughout Europe [36]. Importantly, the female flowers (which are cone-shaped) harbour bitter acids and floral aromas that famously give beer it's 'bitter" taste. These flowers are universally used in modern-day beers, integrating a wide range of additional flavours compare to "hopless" beer [36]. It is therefore no surprise why the integration of hops in beer brewing in Medieval Europe become so popular.

The first mentioning of hops in beer is found in written laws in the St. Peter and St. Stephen monasteries at Corbie, France, during the reign of King Louis the Pious, when it appears to have already been a routine practice [30] Soon after, various monasteries in France quickly adapted hops into their own brewing recipes, and although the mentioning of this practice appears in Germany a few decades later, it is likely German brewers were already using hops, given the existence of "hop-gardens" in Hallertau, Germany, around 1,300 years ago [37].

Finally, a new approach for beer brewing was invented, ultimately sealing the two major classes of beers that we see today. Historically, beer brewing was carried out by yeast that would float to the top after fermentation, thus known as top-fermenting yeast, or *ale* yeast [2]. But around 700 years ago in Bavaria, Germany, a new species of yeast was discovered that instead sunk to the bottom after fermentation, thereby known as bottom-fermenting yeast, or *lager* yeast [2, 38]. Interestingly, lager yeast can ferment at much colder temperatures, between 5-15C, as opposed to ale yeast which required warmer temperatures between 17-22C [2, 38]. As such, the colder temperatures allowed lager beers to last throughout the winters. Lager yeast are also much more genetically complex than their ale counterparts, but I will discuss this in the later chapters.

The popularity of lager beers in Germany found its way to various breweries throughout world, especially in 19th century [2]. As such, there are two major classes of beers that

we see today: ale and lagers, discretized by the type of yeast used during brewing (top or bottom-fermenting). Sure, beer recipes have changed in the past few hundred years, evident by the various Indian Pale Ales, Porters, Stouts, Ambers, Triples, Sours, Hefeweizens, and Pilsners. But in the end, the nature of their brewing is centered by the capabilities—or more accurately, biology—of the yeast used.

0.3 Yeast: man's best microbial friend

Universally, dogs are known as man's *best friend* due to their historical support through hunting, guarding, civil duties, and companionship. But if the past two sections have taught us anything, is that there is a clear contender to this title: yeasts.

As already discussed, yeasts (and their alcohol-fermenting capabilities) have played major roles throughout the development of human civilization, and arguably, shaped the evolution of our species. But despite the long, complex history of their use to make alcohol and bread, yeasts were never really seen as living organisms until much recently.

Ancient brewers knew that yeasts were a critical ingredient in the fermentation process. For example, around the emergence of lager-beer brewing in the 14th century, *Hefners* (or Yeasters) in Nuremberg, Germany, were responsible for harvesting and stocking yeast [2]. In Olaus Magnus', *History of the Northern Peoples*, written in 1555 on Swedish culture, Olaus refers to the practice of *re-pitching*, that is, recycling yeast from previous beer brewing batches to brew the next one [39]. However, it wasn't until the 17th century when the idea of *living microbial organisms* was first scientifically proposed.

In 1674, Antoni van Leeuwenhoek—a Dutchman from the city of Delft, The Netherlands—began to peak at the microscopic world through an early version of a microscope [40, 41]. Initially, it was a simple idea: craft together a glass-lense on a metal apparatus, and use it to amplify objects on the other side [40–42]. Fairly, his microscopes were not entirely novel as several scientists before him had similar ideas, including Hans Lippershey, Hans Janssen, and Zacharias Janssen, who are credited for the invention and use of the first microscope [42]. More knowingly, Robert Hooke published his famous work, *Micrographia*, nine years before, where he described his observations of various objects up-close such as a needle, a flea, and various seeds using his self-crafted microscopes [43]. Likely, Leeuwenhoek was inspired from the work of these individuals, especially from Hooke [41]. However, Leeuwenhoek's microscopes had one big advantage: their magnification capabilities were immensely powerful, even in today standards.

We now know that Leeuwenhoek's microscopes magnified from 30-200x [44, 45], enough to distinguish individual structures of 0.7μm in thickness [45] (to put in perspective, the width of a strand of a human hair is 17-181μm in thickness). The magnification power of his microscopes thus enabled him to make an important observation: there was an entire microscopic world filled with organisms which he called, *little animals* [46]. These *little animals* were found in a wide range of substances including canal and rain water. Importantly, he was the first to describe yeasts under the microscope, regarding them as small "globules" [47]. His observations were first met with great skepticism, especially from Robert Hooke who claimed that he could not recreate his observations [41] (likely due the lack of magnification power obtained in his microscopes). Although Hooke ultimately was able to validate some of Leeuwenhoek's findings [41], it took nearly 200 years to solve the nature of these *little animals*.

Independent observations by Theodor Schwann and Charles Cagniard de la Tour in the late 1830s showed that yeast were actually living organisms, acknowledging observations by Leeuwenhoek nearly two centuries later [48, 49]. Although yeast were known to be a key ingredient in alcohol fermentation, fermentation was thought to be driven purely by a chemical process through the reaction of oxygen and decaying matter, which was heavily supported by the German chemist, Justus Leibig [50]. Until the observations by Schwann and Cagniard, yeast were regarded as either natural by-products or catalyst of the chemical reaction [50].

Both Schwann and Cagniard—with the aid of more advanced microscopes—observed that yeast were actually, "small spherical or oval globules" that decomposed sugars into alcohol [48, 49, 51]. As they appeared to reproduce, they "were not merely a simple chemical or organic substance", proposing instead that yeast were living organisms [48, 49, 51]. Liebig quickly responded with satirical illustration of yeast reproducing and converting sugar to alcohol [50]—ironically, this illustration was probably the most accurate "model" of alcohol fermentation at the time. He instead proposed that yeast were merely decomposing and were reacting with oxygen [50].

Famously, Louis Pasteur showed that yeast reproduced and created alcohol even without the presence of oxygen nor the organic compounds that Liebig suggested were decomposing into alcohol [52]. Subsequent work by Pasteur and Robert Koch ultimately highlighted role of microbes in infectious diseases including cholera, anthrax, and rabies [53]. But, above all else, they showed that microbial organisms—whether friend or foe—could be controlled, exemplified by their pioneering work on acquired immunity via vaccines [51].

What started out as an international debate about the exact role of yeast during alcohol fermentation, ultimately led the field of *microbiology*, revolutionizing medicine, food production, and human well-being. For example, throughout the past century, there has been large investments in studying microbes that may benefit our society. Aside from alcohol fermentation via yeast, a large part of these investments has aimed at harvesting enzymes, which are small, organic compounds that perform specific chemical reactions. Enzymes are produced by most organisms and are generally adapted to the environment that the host organism lives in. As such, enzymes enable microbial organisms to live in diverse environments while making use of different nutritional sources. For example, the bacteria, *Deinococcus radiodurans*, can withstand the highest dose of radiation known to date at 5,000 Gys radiation (to put it in perspective, 5 Gys of radiation is considered lethal to humans), thanks in part to various specialized enzymes that repair damages induced by high radiation [54]. Another example is the bacteria, *Nitrosomonas europaea*, with the unique ability to use ammonia as its main energy source by chemically oxidizing it to nitrate (which is also an important step required for sanitizing waste water) [55].

As such, microbes are generally "screened" for enzymes that may have beneficial chemical properties in industrial applications. *Proteases*, for example, are enzymes that can break proteins into smaller pieces. Protease from a bacteria, *Bacillus licheniformis*, were used to remove residue stains in fabric in the first protease-containing laundry detergent in 1956 [56]. Since then, proteases are key ingredients in standard laundry detergents [57]. Alternatively, proteases from a similar bacterial species, *Bacillus subtilis*, have been harvested to digest proteins in milk to help create "curd effect" during cheese production

[56, 58]. Although proteases make up of more than 60% of the global enzyme market, there are also a variety of others enzymes such as *lipases* (breaks down fats; an industrial example is cocoa butter production), *cellulases* (breaks down plant matter; olive oil extraction), and *lactase* (breaks down sugars in milk; supplemented to people who are lactose-intolerant), all which have been derived from microbial organisms. But it's not only alcohol and enzymes that make microbial organisms interesting to study.

Understanding the medical implications of microbes have vastly improved our personal health and wellbeing. One example was the discovery of antibiotics in the early 20th century. As it turns out, microbes constantly fight with each other for space and resources [59]. One effective tactic that they use is to produce chemical compounds called, antibiotics, which can be toxic to microbes as they can disrupt essential functions necessary for a microbe to live [59]. By producing antibiotics and dispersing them in the surrounding environment, a microbe can effectively reduce nearby competition [59], similar to how yeast can reduce competition to nearby resources with alcohol fermentation. Sir Alexander Fleming—a WWI physician who later became heavily interested in microbiology—studied a species of mould called *Penicillium* in the late 1920's [60]. These species of mould can be found in a variety of damp environments like soil, and commonly cause food spoilage [61]. As noted by several scientist before him, species of Penicillium were known to have antimicrobial properties, that is, able to kill microbes in the surrounding environment [62, 63]. However, Fleming managed to formally describe the antibiotic produced by this mould, called *penicillin*, and proposed its potential use to treat a variety of infectious diseases [60]. Particularly, he showed that they were effective in killing the microbes responsible for causing anthrax, cholera, diphtheria, and typhoid [60] which were common at the time.

With the help of Howard Florey and Ernest Chain from Oxford University, they were able to develop a method to mass produce penicillin [62], and it was introduce in the 1940's to treat wounded soldiers during WWII, ultimately sparking the "Era of Antibiotics" [64]. Although it is difficult to estimate the total number of lives saved by penicillin alone, Allied powers knew about the strategic importance of this antibiotic [65], saving between 12-15% of Allied forces in WWII [66]. In the USA alone, the government invested in 171 different companies to mass produce penicillin [67].

The city of Delft, The Netherlands (hometown of Antoni van Leeuwenhoek), also played an important role in the production of penicillin. The company *Nederlandsche Gist en Spiritusfabriek*, or NG&SF for short, secretly produced penicillin amid occupation of German forces during the 1940's [68]. This was largely due to the Dutch biodiversity institution, Centraalbureau voor Schimmelcultures, or CBS for short, which maintained the largest collection of fungal samples (interesting fact: samples from this institution often contain the prefix "CBS" in their names, which is the origin of microbe studied in chapter 4). Alexander Fleming sent the penicillin-producing sample of mould to CBS [68], but German forces (who occupied The Netherlands at the time) also knew about the strategic importance of penicillin [65] and demanded a sample of the mould [68]. In response, CBS purposely sent the wrong fungal sample to avoid penicillin-use by German forces. NG&SF thus avoided suspicion by constantly offering gin to the local appointed German officers while producing penicillin underground in milk bottles [68].

Since then, various different antibiotics were discovered, especially during the 1950-

1970s which is widely deemed as the "Golden Era of Antibiotics" [63]. And today, we can effectively treat various infectious diseases that would've been regarded as "deadly" 100 years ago [63]. Unfortunately, microbial pathogens are beginning to acquire resistance to our antibiotics, making some infections challenging to treat. Indeed, acquired drug-resistance in microbes is seen as one of the major challenges for society in the next coming decades [69].

Now, this story started out with yeast and its ability to produce alcohol. And despite our long-lived recreational love for both, they continue to be headline-news in microbial research. It wasn't until the late 1990s when an international community of more than 600 scientists from around the world came together to determine the complete genomic sequence of *S. cerevisiae*, a first for eukaryotic organisms [70]. The study revealed a genome of 12 million DNA letters (or nucleotides) containing about 6,000 genes spread out across 16 chromosomes [70]. Importantly, unlocking the genetic code of *S. cerevisiae* had a powerful implications: could we re-write its genome and genetically engineer its capabilities for the benefit of human society?

Throughout the past two-decades, researchers have been able to use yeast as "biological swiss-army knifes", engineering them for wide variety of important industrial applications. In the fight against climate change, various scientists are attempting to engineer yeast with specialized biological pathways in order to breakdown renewable plant biomass and produce biofuels as alternative energy sources [71]. In the promise for cheaper and safer therapeutics, researchers have engineered the complete biological pathway in yeast to naturally to produce opioids, which are commonly used for pain management in (human) patients [72]. Similarly, the complete biological pathway for producing cannabinoids—which are also used for pain management—has also been engineered in yeast [72].

Excitingly, with the rapid progression of genome sequencing technology, we are only beginning to understand the vast genetic diversity of yeast throughout the world [73]. Coupled with promising and integrative genetic engineering technologies, such as CRISPR [74], the engineering capabilities and general strives that yeast can provide to the scientific community is undoubtedly powerful.

1

Introduction

Yeasts have played an influential roles in human history, shaping our societies, sciences, and (bio)technological capabilities[1]. Deservedly, the genome of *Saccharomyces cerevisiae*—commonly referred to as baker's yeast—was the first eukaryotic genome to be sequenced and assembled, thanks to an international consortium of more than X's institutions throughout the 1990s. But as scientists have quickly learned, genomes from individual members in a species are not identical, especially in yeasts.

Yeasts have undergone a complex evolution, thanks in part to human domestication. This is particularly evident in industrial yeasts, which were not only been subjected to external environmental pressures (such as those found in fermentation systems and bioreactors), but often (purposely) "mixed" with other yeast populations. As such, industrial yeasts are often aneuploid (e.g. multiples copies of the individual chromosomes, not always with the same number) and hybrids (chromosomes from different yeast species in the same nucleus). The genome of a single yeast is thus hardly a representation of the true genomic landscape that exists in its species.

To add to the complexity, traditional bioinformatics algorithms don't farewell when analysing aneuploid and hybrid genomes. Aneuploidy is a known hallmark challenge in *de novo* assembly that leads to fragmented genomes, especially those with high sequence variation. And the hybrid-nature challenges our ability to trace their complete evolutionary histories. Although recent progression in long-read sequencing technologies provides various opportunities to overcome these challenges, the data alone is not enough.

This introduction gives an algorithmic overview of fundamental bioinformatics methods surrounding sequence analysis, *de novo* genome assembly, and comparative genomics. It then transitions to the modern methods aiming to leverage both short and long-read sequencing data—inspirations to the contributions of this thesis. Finally, it concludes with an overview of the contents presented in the rest of this thesis and the bioinformatic challenges addressed. An overview of this introduction is shown in Figure 1.1.

[1]There is a whole story about this in the Prelude if you haven't read it yet

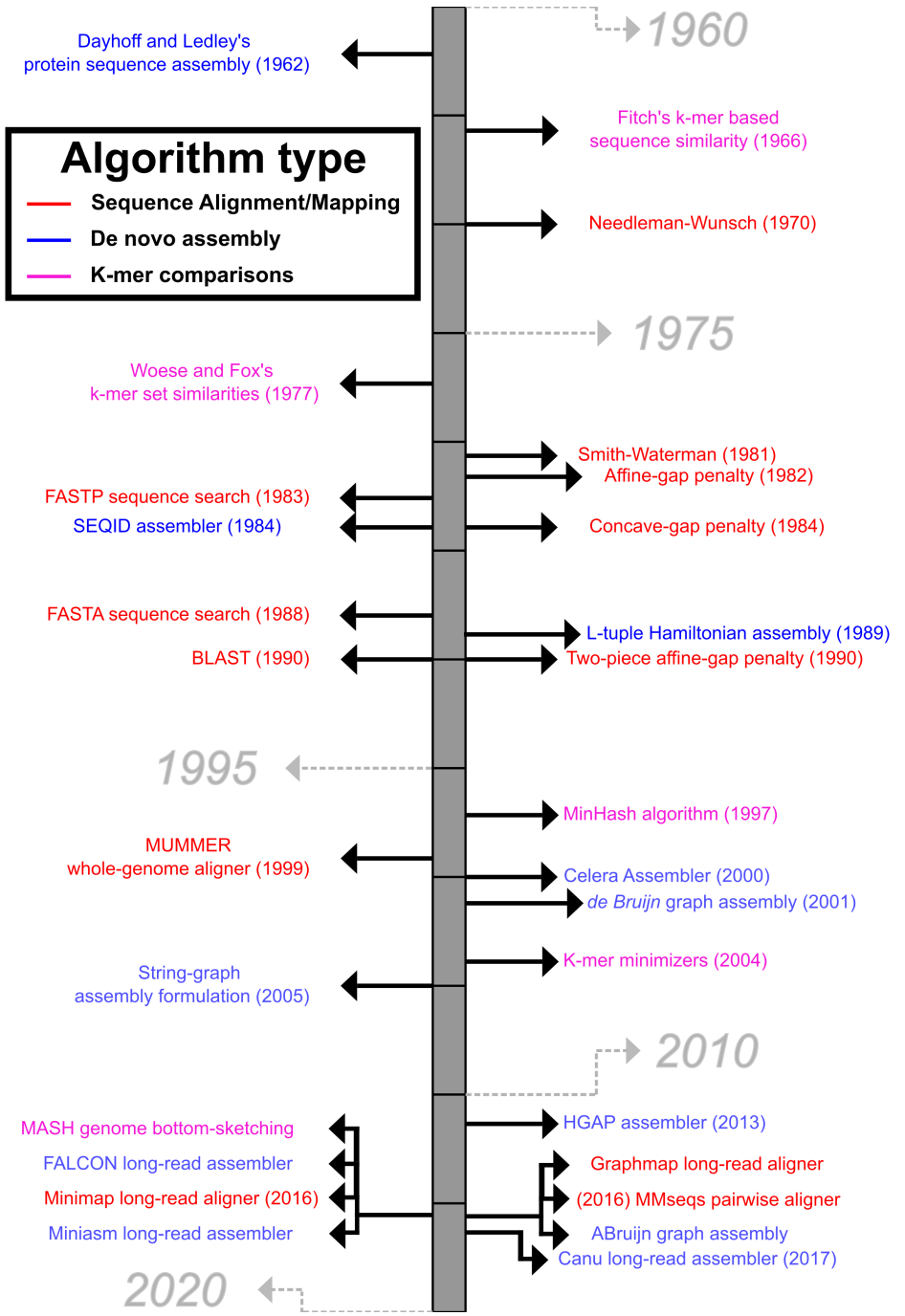


Figure 1.1: Overview of methods and algorithms described in this introduction.

1.1 In the era of long-read genomic data

Today, we have amassed something that many researchers in the 1950s were only beginning to imagine [75]: genome sequencing data. More accurately, A LOT of sequencing data. In 2015 (at the start of my PhD), the European Bioinformatic Institute housed nearly 10^{15} bytes of genomic data (10^{12} bytes which were microbial) [76]. To put in perspective, that's a total of 1.36 million meters of DNA if you were to construct a physical chain it into one giant chemical chain. In fact, genomic data is regarded as a "four-headed-beast" as data generation, storage, accessibility, and analysis rivals that to other "big data" industries, such as NASA, YouTube, and Twitter [77].

This explosion of genomic data is due to rapid progression of genome sequencing technologies. The first generation of sequencing technologies (although slow and expensive) helped generate the first set of "complete" or "early-drafts" genomes, such as bacteriophage ϕX174 in 1977 [78], *S. cerevisiae* yeast in 1996 [70], and the human genome in 2001 [79]. These first genomes provided an invaluable genetic platform that helped researchers better understand the role of genetics in phenotypes (e.g. physical attributes and diseases), whether it be humans, livestock, or microbes. The introduction of next-generation sequencing technology (first demonstrated in mid-2000s [80]) enabled researchers to sequence hundreds to thousands of genomes of any organism at a much cheaper price, unraveling unprecedented information about genomic diversity. More recently, the introduction of third-generation sequencing technology overcomes limitations of its predecessors by decoding larger molecules of DNA, enabling analysis of longer, contiguous stretches of genomic information [81]. Some of the main foci of these technologies have therefore aimed at generating "complete" genomes reconstructions for organisms that were previously studied or recently discovered [81].

Due to technological limitations, sequencing technologies have only been able to decipher or sequence (small) substrings of a chromosome at a time per genome, often equating to algorithmic challenges when analysing genomic data. To better understand these algorithmic challenges, I provide some general notations:

A DNA sequence can be represented a string, s, composed of four nucleotides, $\Sigma = \{A, C, G, T\}$, whose size is denoted by $|s|$. Equivalently, a protein can also be represented as a string but with 22 amino acids [82–84]:

$$\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, O, P, Q, R, S, T, U, V, W, Y\}$$

If a single chromosome can be represented as a single string, than the genome of an organism can be represented as a set of strings, $S = \{s_1, s_2, s_3, \cdots, s_n\}$, where n, or equivalently the set size $|S|$, represents the total number of chromosomes, and the sum of the sizes of all chromosomes, $Size(S) = \sum_{i=1}^{n} |s_i|$ is the genome size. For example, a human a genome with 23 chromosomes, $|S| = 23$, has a genome size of $Size(S) = 3.2$ billion nucleotides. Similarly, for yeast, $|S| = 16$ chromosomes and a genome size of $Size(S) = 12$ million nucleotides.

Ultimately, genome sequencing technologies generate a set of reads, $R = \{r_1, r_2, r_3, \cdots, r_m\}$, where each r_i is a substring from some chromosome in S with a *sequencing error rate*, ϵ. Generally, $|r_i| << Size(S)$ even with recent technological innovations. As such, the algorithmic challenge in analysing genomic data has thus largely centered in using R to reconstruct (an approximation of) the original genome(s), its gene-contents, and its evolutionary relationship to other individuals or organisms.

1

The growing complexity of the four-headed-beast has required sophisticated computational techniques in order to efficient manage and analyze genomic datasets. Consequently, bioinformatics has rapidly evolved throughout the past few decades, adapting to the progression, limitations, and ambitions of the genomic data produced by sequencing technologies. Interestingly, although there is a plethora of bioinformatics methods published every year—particularly those revolving sequence analysis—many describe techniques based on prior established methods (as expected as they directly draw inspiration from them), and some cases, can be regarded as "digital versions" of experimental techniques published several decades before.

1.1.1 On the fundamentals of sequence alignment

Shortly after the sequencing of the first set of proteins in the 1950s researchers realized that comparing protein and/or genomic sequences among different individuals or organisms could provide insights about genetic diseases and evolutionary histories. It is therefore not surprising that a major theme in bioinformatic research in the past 70 years has focused on *sequence alignment*, an algorithmic scheme to comparing DNA or protein strings. It is important to distinguish common terminology in this field: when I refer to sequence alignment I am referring the algorithmic procedure to comparing two strings; not to be confused with *sequence mapping* which I refer to as the algorithmic procedure to identify the location of substring among a set of much larger strings (e.g. a set of chromosomes representing a genome). Importantly, the introduction of long-read sequencing data has required new methods for sequence alignment and mapping, all which use and extend established methods first developed in the last quarter of the 20th century.

Score optimization via dynamic programming has traditionally been the main fundamental approach to sequence alignment. Starting with the Needleman-Wunsch algorithm for global sequence alignment in 1970 [85], and the Smith-Waterman algorithm for local sequence alignment in 1981 [86], these two methods have served as the core functionality in most sequence mappers and aligners in the past few decades. And although the dynamic programming paradigm for sequence alignment is consistently used by various methods, they have been refined by a variety of techniques and timely-problems, such as handling structural sequence variation which has resurfaced as a major theme genomic research in the past few years.

Initially, sequence alignment used a single-gap value for both global and local sequence alignment [85, 86]. For example, given two strings, a and b, the dynamic programming algorithm for aligning both strings globally and locally can be seen as finding a path in a scoring matrix, M, that minimizes some score. Specifically, a m x n scoring matrix (where m and n correspond to the size of the two strings) is first initialized such that $M_{(i,0)} = M_{(0,j)} = 0$. The scoring matrix is then filled through the following approach:

$$Global : M_{(i,j)} = max \begin{cases} M_{(i-1,j-1)} + score(a_i, b_j), \\ M_{(i-1,j)} - p, \\ M_{(i,j-1)} - p \end{cases} \tag{1.1}$$

$$Local: M_{(i,j)} = max \begin{cases} M_{(i-1,j-1)} + score(a_i, b_j), \\ M_{(i-1,j)} - p, \\ M_{(i,j-1)} - p, \\ 0 \end{cases} \tag{1.2}$$

Where $score(a_i, b_j)$ is a pre-defined constant integer whenever there is a matching nucleotide, $a_i = b_j$, or a mismatching nucleotide, $a_i \neq b_j$; and p is the gap-penalty penalizing insertion and deletions between the two strings, as $M_{(i-1,j)}$ or $M_{(i,j-1)}$. In both the global and local alignment scheme, both m and p are constant through the entire alignment. However, it became clear that this design can create ambiguity when representing and identifying structural variation.

Osamu Gotoh in 1982 [87] proposed the affine gap penalty technique to yield the alignment in the left (e.g. a single deletion event of two nucleotides more likely than a two different deletion events). Specifically, the cost of a gap takes the form $p = ek + o$, where e is the cost of extending it, k is the number of gaps introduced, and o is the cost of opening a gap; effectively reducing the time complexity from $O(a^2b)$ to $O(ab)$ in comparison to previous methods integrating multiple gaps. However, there can be multiple optimal alignments and Gotoh is only guaranteed to find one as it only follows one of the potential multiple paths in alignment matrix. As such, Stephen Altschul and Bruce Erickson in 1986 [88] provided and improved approach of the affine gap cost that enables the identification of multiple optimal alignments while remaining $O(ab)$. This was possible by using three arrays representing the different possible paths an alignment can take: diagonally (as a match) and horizontally and vertically, each representing a gap extension or gap opening in respects to both sequence. By tracking the direction of each path, all optimal alignments can be identified through joint traceback of the three matrices. This can be represented as:

$$\begin{aligned} M_{(i,j)} &= max \begin{cases} M_{(i-1,j-1)} + score(a_i, b_j), \\ A_{(i,j)}, \\ B_{(i,j)} \end{cases} \\ A_{(i,j)} &= max \begin{cases} A_{(i-1,j)} - k, \\ A_{(i-1,j)} - (e+o) \end{cases} \\ B_{(i,j)} &= max \begin{cases} B_{(i,j-1)} - k, \\ B_{(i,j-1)} - (e+o) \end{cases} \end{aligned} \tag{1.3}$$

Later, Gotoh in 1990 [89] further improved the general goal of aligning sequences with large structural variation by employing a two-piece affine gap penalty using a similar technique to that of Altschul and Erickson to retain the algorithm in $O(ab)$ while also identifying all possible optimal alignments (described in more detail in section 1.1.3).

As will become more clearer in the later sections, long-read sequencing data enables investigation of structural changes in a genome, such as large deletions and insertions. However, they are plagued by high ϵ often comprised of smaller insertion and deletion events. Adaptations of the single and two-piece affine gap penalties in modern sequence

1

mappers have thus helped correct sequencing errors [90–92] and identifying true large structural events in a genome [93, 94] (see section 1.1.3).

Other gap-scoring functions have been investigated, such as the concave gap-function which takes the form, $p = eln(k)+o$, first introduced by Michael Waterman in 1984 [95] and later optimized by Webb Miller and Eugene Myers in 1988 [96]. The rationale being the size of indels were observed to exponentially grow, and thus needed to be modelled logarithmically [95–97].

In parallel to the research above, researchers began to focus in optimizing the speed of sequence alignment as enomic databases began to grow. For example, as opposed to having a high resolution between the differences of two strings, the focus instead became identifying *target* strings that have high similarity to a given *query* string (e.g. aligning a newly sequenced gene across many existing genes). More specifically, similarity of two strings, $Sim(query, target)$, measures the proportion of matching bases to total bases in the alignment of query and target (this exact calculation varies through different scoring schemes). John Wilbur and David Lipman in 1982 [98] proposed an efficient method to search some given query string across a database of many target strings of variable lengths. They highlighted the observation that the alignment of two strings high similarities will harbour high-scoring alignment paths in the scoring matrix M. However, identifying the optimal alignment via a dynamic programming scheme is computationally expensive and significantly slows down any searching algorithm that directly employs this technique.

Instead, Wilbur and Lipman introduced the concept of $kmer$ matches (or k-tuples as presented in the paper [99]) as a mean to quickly compute an approximate global-alignment between the query and target strings without needing to compute the expensive alignment scoring matrix. The set of k-sized kmers for a string s of size n can be regarded as all k-sized prefixes of s. If $subs(s, 0, 3)$ yields the first three characters of s, then $Kmers(s,k) = \{subs(s,0,k), subs(s,1,1+k), subs(s,2,2+k), ..., subs(s,n-k,n)\} = \{k_1, k_2, k_3, ..., k_(n-k+1)\}$. Wilbur and Lipman utilize a single one-dimensional array for a given query string of length $n-k$ where each element in the array represents a kmer in the query storing a list of starting positions of that kmer in the query sequence [99]. For a given target sequence, the respective location of kmer matches—that is, kmers in both $Kmers(query, k)$ and $Kmers(target, k)$—can be identified using the one-dimensional array. The array can then be used to identify and cluster consecutive kmer matches, which would appear as a high-scoring local alignment paths in the scoring matrix. Within each cluster of kmer matches, they infer sequence similarity by positively scoring the number of matching ordered kmers penalized by size differences between kmer matches [99]. The similarity of the query and target strings can thus be approximated based on the number and score of high-scoring paths. As discussed in the later sections, k-mer matching (also known as k-mer seeding) became an imperative step in long-read (and whole-genome) mapping and alignment, as well as probabilistic "alignment-free" methods, later discussed in section 1.1.5.

Wilbur and Lipman subsequently improved their method with the introduction of FASTP in 1985 [98] which was meant to rapidly compare protein sequences to continuously growing databases such as the NBRF protein database. Using a similar scheme as their method in 1982, they improved the high-scoring diagonal search by integrating amino acid substitution matrix (e.g. PAM250) to increase sensitivity of finding homology

between distant proteins sequences. In other words, as opposed to having a constant positive score for any matching charactering the $M_{(i,j)}$, the score depends on the corresponding value in the given substitution matrix specifying how the two amino acids should be scored based on the likelihood of the two mutating into each other. Additionally, they also introduced the use of a "banded" Needleman-Wunsch algorithm to compute an alignment only for a narrow path in the scoring matrix by restricting the values of i and j within a proximity of the high scoring diagonal [98]. They further improved FASTP by introducing FASTA [100] which not only improved sensitivity when searching for sequences in a database, but also enabled queries of DNA sequences. The major difference is that FASTA additionally enables the calculation of an approximate global alignment by linking high scoring diagonals and computing a banded global alignment restricted through linked diagonals.

In 1990, Stephen Altschul along with David Lipman, Warren Gish, Webb Miller, and Eugene Myers proposed an improved search method, BLAST [101], was based on prior work of the FASTA method through three main steps. Similar to FASTA/FASTP, the first step aimed at extracting kmers and their high scoring neighbours, that is, a kmer a long with variants of that kmer that differ by no more than d nucleotides, increasing the sensitivity of identifying kmer match between the two strings. For each kmer match, it is extended in both directions, positively scoring matches and penalizing mismatches, up until a (local-)maxima is reached (referred to as maximal scoring pairs). It's analogous to finding high-scoring diagonals as done in FASTP/FASTA, but it is based on the extension of individual kmer matches as opposed to first identifying high scoring clusters of kmer matches and linking/extending those to other clusters [101]. Importantly, base-level extension uses a X-drop metric to stop an extension at the point where the score for $M_{(i,j)}$ drops below some threshold [101]. This is particularly useful when two sequences share common substrings, but their respective ends are not similar. As we will see in the next few sections, the fundamentals of these techniques are used to not only rapidly map and align long-read sequencing data, but to efficiently compare *pan-genomes* (e.g. the gene content of microbial populations), which have exponentially increased throughout the past decade.

At the turn of the new millennium, the "complete" whole genomes were being constructed, including human and various microbes such as the yeast *Saccharomyces cerevisiae* [70] and the bacterial pathogen *Staphylococcus aureus* [102]. As such, there were efforts to develop methods that would enable whole-genome alignments, as opposed to local-sequence alignment across large sequence databases. More specifically, whole-genome alignment can be seen a form of semi-global sequence alignment for the entire genome, but it becomes increasingly challenging when we consider structural variations and homologous sequences, which are pronounced features in many (microbial) genomes. A detailed review of whole-genome alignment methods has been recently published [103]. However, to better understand the downstream computational challenges that arose from long-read sequencing more than a decade later, I provide quick summary of two specific methods during early development of whole-genome aligners: kmer matching and chaining.

The same logic utilized by FASTA/FASTP and BLAST applies when performing whole-genome alignment: it is much more efficient to identify and compute sequence alignments through kmer matches, often referred to as *anchors* or seeds, as opposed through a brute-

1

force search across the entire genome. This is particularly important when we consider that two genomes may have shared evolutionary sequences but are located in different regions in their respective genomes due to evolutionary events such as horizontal gene transfer.

MUMmer was one the first whole-genome aligners and is still widely used today [104]. The underlining key to this method is the use of *suffix trees*, which represent all suffixes in a genome and their locations in a tree data structure. Unlike kmers, suffixes in the suffix tree are not restricted to a specific size, but can vary in length. By constructing a suffix tree for two given genomes, maximal unique matches (MUMs)—longest possible sub-sequences that are shared by both genomes and occur only once—can be quickly identified and serve as anchors for potential alignments between the two genomes [104]. The $MUMs = \{k_1, k_2, k_3, \cdots, k_n\}$ are then subjected to a chaining algorithm to find collinear regions between two genomes [104]. In MUMmer, the chains are computed via the longest increasing sub-sequence algorithm, such that the starting positions of each $kmer_s tart(k_i)$ in the two genomes are respectively ordered, $start(k_1) < start(k_2) < start(k_3), \cdots, < start(k_n)$, leading to collinear regions that can later undergo more sensitive sequence alignment to an "extend" step at each anchor [104]. In fact, chaining of biological sequences was already being discussed by David Sankoff in 1972 [105]. And as described in much more detail in the next section, recent (long-read) aligners propose different approaches for computing alignment chains, such a heuristic-scoring metrics based on the coverage of the reference [93, 103], and formulation as the *0-1 knapsack* problem [94].

The purpose of suffix trees is effectively the same as kmer hashing, an alternative approach to finding matching kmers or anchors. In short, a hash function applied to a string yields an integer value regarded as the hash value, $H(s) = h$, where h is uniformly distributed between a fixed range based on the corresponding computer architecture (e.g. 32 or 64 bits). Applying $H(s)$ across all $Kmers(s, k)$ in a sequence or genome yields a collection of hash-values. Matching kmers between two genomes or sequences can thus be identified via the intersection of integers between the two collections. As we will also see in the next few sections, various techniques have been developed to ease the application hash functions on genomic sequences, such as the $(w - k) - minimizers$ proposed by Michael Roberts et al. which only retains the "smallest" kmer of length k in a window size w [106], and recent dimensionality-reduction techniques to significantly reduce the total number of kmers required to store and compare [107]. Additionally, hash representation of sequences and genomes is widely used in long-read mapping and alignment, as well as recent genomic streaming algorithms aiming quickly compare thousands of (microbial) genomes (see section 1.1.5).

Overall, the strategy of anchor-finding and chaining is employed in various other whole-genome aligners, although each method uses different strategies that compute and refine anchors in order to address homologous sequences and structural variation [103]. As previously mentioned, this is in essence the same logic used by sequence aligners in the 1990s, such as FASTA/FASTP and BLAST. And not surprisingly, it is also the same logic employed by sequence aligners aiming to make use of sequencing data produced by the so-called "next-generation sequencing data" [81, 108]. To be fair, the problem was much different when introduced in the mid 2000s, since the goal of aligning two (complete) whole genomes differs to that of aligning millions of "short-reads" (e.g. less than

a few hundred nucleotides) to a single high-quality "reference" genome [108]. But it is still a problem of sequence mapping and alignment. Uniquely, the sheer magnitude of the data generated by these new sequencing technologies required new compression and indexing techniques, such as the Burrows-Wheeler transform and FM-index to quickly and efficiently map and align millions of reads [108]. Similarly, there are in-depth reviews of sequence mappers and aligners for next-generation sequencing data [108].

Nevertheless, when third-generation sequencing was introduced, the relatively large read-lengths, high-error rates, and high-throughput of the data generated required a new generation of sequence mappers and aligners, making use of ideas proposed throughout the several decades before.

1.1.2 *De novo* genome assembly: the early days

In parallel to research in sequence alignment, there were also major contributions to the problem of *de novo* genome assembly. Recall that sequencing technologies output a set of reads $R = \{r_1, r_2, r_3, \cdots, r_m\}$, with some error rate ϵ and average read-length l representing a collection of substrings from a genome of one or more chromosomes (or strings), $S = \{s_1, s_2, s_3, \cdots, s_n\}$. More specifically, given a high concentration of the same DNA molecule(s) (e.g. chromosomes), each molecule is sequenced up to some size, typically l. This is because sequencing technologies have been historically limited to producing read-lengths that are significantly smaller than their respective genomes [81, 109], e.g. $l << Size(S)$. However, with enough sequencing data (e.g. high number of DNA molecules), reads will begin to overlap as the number of reads starting from the same position in a chromosome follows a Poisson distribution [110, 111]—the average number of reads per position is referred to as the genome $coverage$. Therefore, in *de novo* genome assembly, the goal is to identify overlapping substrings in R and reconstruct them to an assembly, S', representing (an approximation of) the original chromosomal strings in S. Not surprisingly, *de novo* genome assembly methods have historically reflected the strengths and limitations of sequencing technologies of their times. But from an algorithmic perspective, they have largely concentrated on two main approaches: overlap-layout-consensus and *de Bruijn graphs*.

The first *de novo* genome assemblers were proposed in 1979-1980, in conjunction with the first set of whole-genome sequencing data. As demonstrated by Fred Sanger via the *shotgun sequencing* method [112], early whole-genome sequencing data yielded reads of 200-300 nucleotides [112], formatted in gel-based pictures. As a result, Thomas Gingeras et al. [113] and Rodger Staden [114, 115] proposed the first algorithms which identified read-overlaps through exact substring matching on these gel-based pictures. The overlaps were ordered and merged, generating a contiguous consensus sequences [116], termed *contigs*[115]. In essence, these were the first set of algorithms following the overlap-layout-consensus paradigm which is one of the two main paradigms used today in *de novo* assembly algorithms.

More rigorous, came the algorithms proposed by Hannu Peltolta et al. [117] and John Kacecioglu and Eugene Myers [118] which generalized the overlap-layout-consensus approach [116]. In short, *de novo* genome assembly was viewed as the shortest common superstring problem, aiming to identify the shortest possible string, S', that can be reconstructed from R such that each read aligns to substring of S' with an error rate of ϵ [116].

1

This problem could be solved by constructing an overlap graph [119, 120], $G = (V, E)$, where E is a set of edges created between two reads, r_1 and r_2, if the suffix of r_1 overlapped with the prefix of r_2, denoted as $r_1 \longrightarrow r_2$, with the overlap size denoted as length $length(r_1 \longrightarrow r_2)$. The vertex set V is thus the unaligned substrings in each overlap (e.g. the unaligned suffix of r_1 and unaligned prefix of r_2). In theory, a ranking of the edges by their size and computing a *Hamiltonian path*—a path in the graph that visits a node exactly once—could solve the shortest common substring problem, where contigs could be inferred by the concatenation of the initial suffix and downstream prefixes of the path, followed by a consensus sequence generation to correct for sequencing errors [116]. However, technical and biological challenges made this problem much more difficult.

One challenge is to accurately construct the overlap graph. An overlap graph could be accurately constructed from whole-genome sequencing data by constraining the graph to pairs of reads where the probability of overlapping by chance was significantly small [109, 116]. For example, a pairwise sequence alignments of all reads combine with a heuristic filter, could generate edges where $r_1 \longrightarrow r_2$ is unlikely to occur by chance [109, 116]. However, *sequencing errors* (the incorrect interpretation of the DNA sequence by the sequencing technology in a read) could complicate the heuristic assessment of the significance of an overlap [81, 109, 116]. For the first two generations of sequencing technologies, sequencing errors were known ahead of time (early sequencing technologies in the 1980s and 1990s ranged from 2-5% [81, 109, 116]) and could thus be incorporated into the heuristics, such as assuring that respective overlap sequences of each read have $Sim(overlap of r_1, overlap of r_2) \leq 2-5\%$. *Chimeric reads* (resulting from an incorrect physical joining of two non-neighbouring DNA molecules during sequencing preparations) would lead to incorrect biological overlaps [81, 109, 116]. Since the chimeric reads were generally random (e.g. subregions in the genome that were erroneously joined were random), true chimeric overlaps could be identified in the graph as the expected number of reads overlapping a chimeric junction << to the average coverage of true biological junctions (given that the genome coverage is high enough) [121, 122]. As discussed in the later sections, various types of techniques and heuristics are employed to carefully construct the overlap graph, since long-read sequencing data harbour error rates much larger than the first generations of sequencing technologies and their laboratory preparations often yield some fraction of chimeric reads [81, 109].

Ultimately, sequences in genomes are not random, and the most prominent biological challenge (which still plagues all *de novo* genome assemblers today) are genomic repeats [81, 109, 116]. Genomic repeats induce false overlaps between reads, complicating the overlap graph and the generation of (large) contigs. There is a simple solution to resolving repeat-induced overlaps: use only reads that are large enough to span all repeat sequences in the genome and anchor to unique flanking regions in the genome [81, 109, 116]. However, this is an intractable solution as repeats can range from a few tens to several hundreds of thousands of nucleotides (depending on the organism), much less than the even the most recent sequencing technologies [81, 109, 116]. Repeats could still be assembled, but their sequence were often incorrectly compressed to a much smaller size. Some repeat-induced overlaps could be removed from the overlap graph by looking at sequence variation in their overlaps, as true biological repeats were not always exact duplicate copies [109, 116]. In other words, overlaps originating from distinct repeat regions may have lower similar-

ity than overlaps of the same repeat region, which could be used to filter out false overlaps. But this approach largely depends on the divergence of the repeat sequence and error rate of the sequencing technology [109, 116]. Nevertheless, these early methods were sufficient to assemble small microbial genomes, such as viruses (e.g. bacteriophage lambda has a genome 50 Kbp) which often harboured minimal repeat content [81, 109, 116].

The development of paired-end sequencing data in 1990s provided "long-range" information useful in *de novo* assembly [123, 124]. Unlike the traditional shotgun sequencing approach which only sequenced one end of a DNA molecule, paired-end sequencing sequenced both ends [81, 109, 116, 123, 124]. Since the resulting reads originated from the same molecule, one can infer "long-range" information in the assembly, limited to the fragment size of the input DNA molecules. For example, although 10 Kbp DNA fragment yields only two reads each of 700 bp, the fragment may span a large repeat of several thousand nucleotides in size. The overlap-layout-consensus approach will likely lead to two separate contigs as it fails to assemble a repeat much larger than the read-lengths in the read set [81, 109, 116, 123, 124]. But an alignment of the reads to the contigs would show each read aligning to one end of each contig, providing ordering and orientation information of the contigs [81, 109, 116, 123, 124]. As such, the paired-end information was often use to generate an order and oriented version of all contigs in the assembly, often referred to as *scaffolds* [81, 109, 116, 123, 124].

With further developments in sequencing technology including higher sequencing throughput and longer reads (e.g. 1Mbp of sequences with average read-length of 700 bp and error rate of 2% [81, 109, 116]), researchers embarked on sequencing and assembling much larger genomes with higher repeat contents, such as those of yeast, fruit flies, and humans. In conjunction with the *Drosophila*-genome project, *Celera Genomics* (a private company at center of in first human genome sequencing project), which was spearheaded by Eugene Myers, developed an assembler aiming to automate all steps in a *de novo* assembly process [125]. The hierarchical procedure took four steps: *a priori* filtering of repeats, followed by an automated pipeline consisting of pairwise read-alignments, contig assembly, and scaffold generation [125]. Due to the inherent challenges of repetitive sequences, Myers *et al.* exploited known repetitive sequences (e.g. ribosomal DNA) that had been already curated for *Drosophila* from prior genome characterizations by "trimming" substrings in reads that aligned to the repetitive sequences, facilitating an accurate construction of the overlap graph [125]. Consequently, repeat sequences (even collapse versions) could not be assembled. A pairwise alignment of these trimmed reads enabled the construction of the overlap graph, and contigs could be identified via a "unitigger" which identified all maximal paths in the overlap graph without a conflicting edge to another path. The paired-end information of the reads was then used to orient and order contigs into a scaffold [125].

Given the challenges that repetitive sequences imposed on the overlap-layout consensus approach, Eugene Myers alternatively proposed the *string-graph* representation for *de novo* assembly [120, 126] . In this approach, the goal was to construct an overlap graph with consistent genome coverage, as biological and technical challenges often created low and high-coverage sub-graphs (e.g. chimeras, contaminants, and repeats). This so called, *string graph*, could be heuristically achieved by first constructing an overlap graph and performing a *transitive reduction*, that is, the removal of edges that are redundantly rep-

1

resented by a longer path (e.g. the edge $r_1 \longrightarrow r_2$. can be transitively reduced to edge $r_1 \longrightarrow r_3 \longrightarrow r_2$) [120, 126]. The reduction can simplify the overlap graph into sub-graphs with nodes harbouring a single incoming and outgoing edge, and thus compressed into a compound edge. The resulting graph is thus the string graph representation [120, 126]. Repeat sequences would still be compressed into a single compound edge, but a contig could be more accurately generating via a traversal that reflected the local coverage of each compound edge [120, 126].

Importantly, the most computationally demanding step was computing pairwise sequence alignment for all reads in R, even after parallelisation [125]. The number of pairwise alignments is quadratic, requiring $|R|2$ comparisons. And with sequence alignment running at a worst case of $O(MN)$, *de novo* assembly an overlap-layout-consensus approach could become computationally expensive with higher sequencing throughput (e.g. more reads) and longer reads. As we will see, recent *de novo* assemblers for long-read sequencing data use specialized sequence aligners that can quickly approximate $Sim(overlap(r_1, r_2))$, significantly reducing the runtime of an overlap-layout-consensus approach for *de novo* assembly.

However, it was the de Bruijn graph approach—which was first "genomically" discussed by Pavel Pevzner [127]—that would dominate algorithmic efforts in *de novo* genome assembly throughout the past two decades. By the mid-2000's a next generation sequencing (NGS) would begin to fundamentally change biological research, as researchers could now sequence sequence billion of single or paired-end reads in single experiment, enabling much higher throughput (and hence coverage) of genomes, with easier laboratory preparations, lower costs, and lower error rates (e.g. $\epsilon < 1\%$) [81, 109, 128]. A major limitation were its relatively "short-reads", as their read-lengths ranged from an initial 36 bp, to now maximum of 300 [81, 109]. Given the quadratic runtime in the pairwise alignment of the overlap-layout-consensus, and the inherently "small" maximum read-overlap in NGS data, the *de Bruijn* graph assembly approach became widely adopted due it's relatively easier computational resource tractability [109, 116].

In short, in a *de Bruijn* graph, $DB(R, k) = (V, E)$, the vertex set contains all $k-1$ kmers, $Kmers(r_i, k-1)$, for every read, r_i, in R. The edge set encodes all $Kmers(r_i, k)$ and their corresponding edges, e_i, connecting two vertices, $v_a \longrightarrow v_b$, if v_a matches the $k-1$ prefix of e_i and v_b matches the $k-1$ suffix of v_a [116, 129–132]. The *de Bruijn* graph can be thought as a special case of a string-graph under a fixed k-1 sequence overlap [116]. In theory, by "balancing" every node to have the same in/out degree, Euler's theorem can be applied to identify all Eulerian cycle(s), that is, a path where every edge is visited exactly once [116, 129–132], corresponding to contig(s) of the sequenced genome. Computationally, it is more tractable to compute Eularian cycles in a *de Bruijn* graph as it's run-time is roughly proportional to the total number of edges [116, 129–132]. Additionally, *de Bruijn* graphs avoid the costly pairwise alignment step, which often leads to to scalability challenges in the overlap-layout-consensus approach [81, 109, 128].

But much like the overlap and string graph, *de Bruijn* graph assemblers also required heuristics to resolve technical and biological challenges. Similarly, chimeric reads, sequencing errors, and biological contaminants would lead to erroneous nodes and paths such as bubbles and tips [116, 129–132]. With some heuristics, these errors could be removed by filtering edges and nodes with low-coverage [116, 129–132]. The paired-end

information of reads could also be used to identify erroneous paths [110, 123–126]. Importantly, the size of *k*was extremely influential, as *k* needs to be large enough to represent significant overlap of sequences, but small enough to mitigate sequencing errors [116, 129–132]. Inevitably, repeat-induced overlaps still posed major challenges leading to compressed repeat sequences and small contigs [81, 109, 128].

Regardless of the approach, *de novo* genome assembly largely assumed a haploid genome, that is, a single copy for each chromosome. In reality, various organisms—especially higher eukaryotes—are *non-haploids* with two or more copies per chromosome. Genetic variation within multiple copies of the same genome can lead to differences in gene expression and function. As such, the exact sequence per chromosome copy, termed *haplotypes*, can provide insightful information about the genetic basis behind an organism's phenotypes [133] and serve as detailed markers for breeding programs of livestock, improvements of crops, and (industrial) microbial strain engineering [134–136]. Similarly, multiple copies of the same gene can lead to higher expression levels, given that all copies are functional [137]. As discussed in the later chapters, favourable industrial characteristics that make some *Saccharomyces* yeast more robust in producing beer, wines and other alcohols originate in their non-haploid nature [137]. Assembling these types of genomes are therefore industrially important [137, 138].

Unfortunately, non-haplid genomes are challenging to assemble *de novo* with NGS data. In short, heterozygous sequences (which increases the complexity of a *de Bruijn* graph due to an increase of alternate paths) and fluctuations in coverage (which can lead to a shorten a Eularian cycle) generate to more fragmented assemblies [137, 139–141]. And although some heterozygous variation can be identified and characterized, the limited read-lengths of NGS data is insufficient to completely infer haplotypes and proper ordering of contigs into scaffolds [137, 139–141]. Microbial studies—such as those studying *Saccharomyces* yeast—often opt to sequence only haploid strains as it simplifies the *de novo* genome assembly process [142].

Ultimately, the quality of *de novo* genome assemblies from NGS data never really came close to that of established high-quality reference genomes. Which is why genomic projects often used a reference-genome-based strategy: sequence hundreds to thousands of individual genomes via NGS, align reads to an established reference genome, and infer sequence variation by identifying differences between a sequenced genome and the established reference. This strategy dominated the genomics field throughout the past decade. Indeed, computational efforts focused on optimizing algorithms for (reference-genome) sequence mapping and variant calling. But with advancements in third-generation sequencing promising to overcome the limitations of NGS, "old" algorithmic challenges in sequence alignment, *de novo* genome assembly, and whole-genome alignments became a central research topic.

1.1.3 Long-read sequence mapping and alignment

In the past few years, third-generation sequencing has once again resurfaced various algorithmic challenges in sequence alignment, mapping, and *de novo* genome assembly. An immediate advantage of long-reads is its ability to overcome the size limitations of NGS data and generate more complete *de novo* genome assemblies [143–147]. Additionally, long-reads also provide significant advantages when following a reference-genome-based

1

strategy, with their ability to more reliably identify structural variations [81, 148, 149]. In the case of mapping long-reads to a reference, this requires consistent representations of breakpoints to accurately call structural variations [81, 93, 94, 148, 149]. The same problem is similarly seen when mapping and aligning RNA third-generation sequencing data, as mRNA molecules can structurally vary due to exon and intron skipping and retention [93, 94, 150]. For *de novo* assembly applications, pairwise alignments of long-reads are huge bottle-neck due to the computational resources that they demand [151–153]. As such, efficient aligners are needed to perform this crucial step in overlap-layout-consensus for *de novo* assembly, while dealing with deletion and insertions errors that have historically plagued all third-generation sequencing technologies [81]. Ultimately, long-read sequencing has also provided more complete assemblies of (microbial) genomes, motivating us to investigate proteomes and pan-genomes often requiring pairwise alignment of large sets of proteins to identify (novel) gene-families.

Algorithm 1 General long-read sequence alignment

1: **procedure** LONG-READ ALIGNMENT(R,S) ▷ *R*: long-read set; *S*: target-sequence set
2: $K \leftarrow$ *Extract kmers from S using* $\begin{cases} (w-k)-minimizers \\ gapped\ \ q-grams \\ max\ \ frequency\ \ threshold \end{cases}$
3: **for** $r \in R$ **do**
4: $H \leftarrow$ *Identify anchors (e.g. kmer hits)*
5: $C \leftarrow$ *Compute chain(s) using* $\begin{cases} positional\ \ offsets \\ clustering \\ longest-common\ \ substring \\ max-weighted\ \ subset \end{cases}$
6: $A \leftarrow compute\ \ base-level\ \ alignments$
7: **return** A **or** C
8: **end for**
9: **end procedure**

Nevertheless, there have been a variety of different novel or re-purposed methods for sequence mapping and alignment over the past decade to address the computational challenges imposed by third-generation sequencing. In this sub-section, I provide an overview of a subset of these methods, particularly the most recent and (currently) widely adopted ones. More specifically, *Minimap2* [93] and *Graphmap2* [94] employ adaptations of the techniques described in section 1.1.1 to efficiently map and align large datasets of third-generation sequencing technologies, while handling (large) structural variation. NGLMR [149] aims to address heterozygous structural variation in diploid genomes by carefully curating breakpoint signals in long-read alignments. While *MMseqs2* [154] focuses on an indirect consequence of next-generation and third-generation sequencing data: scaling the clustering of homologous genes in thousands of microbial genomes. Despite the different heuristics and approaches that they employ, they can be generalized in Algorithm 1, where lines 4 and 6 correspond to a combination of one or more of the listed procedures.

Minimap2

Minimap2 [93] (a successor to *minimap* [155]) maps and aligns long-reads via traditional kmer-based anchoring, as first done by Wilbur in Lipman in 1982 and later adapted in many whole-genome aligners and second-generation read mappers. However, *minimap2* uses the concept of *minimizers*, the smallest kmers in some defined windows throughout an entire sequence, inspired from Michael Roberts in 2004 [106]. More specifically, minimizers for a collection of kmers, $K = Kmers(s,k)$ for a window size, w, can be computed as:

$$Minimizers(K,w) = \begin{Bmatrix} min\{H'(k_1), H'(k_2), \cdots, H'(k_w)\}, \\ min\{H'(k_2), H'(k_3), \cdots, H'(k_{w+1})\}, \\ \cdots, \\ min\{H'(k_{|K|-w}), H'(k_{|K|-w+1}), \cdots, H'(k_{|K|})\} \end{Bmatrix} \tag{1.4}$$

Where $H'(k_i)$ is either hash value of the smallest lexicographic string between k_i and it's reverse complement, $\overline{k_i}$, or the smallest hash value between k_i and $\overline{k_i}$. The motivation being that storing all kmers in a reference genome could demand high computational resources. Instead, $Minimizer(K,w)$ stores only a fraction of them, and therefore, kmer matches, or anchors, between a long-read and the reference only requires identification of subset of kmers, termed *minimizer-hits*.

The minimizer-hits are then curated by finding an optimal set of candidate anchors that represent an optimal collinear region between the two sequences, a re-current problem in both whole-genome and approximate sequence alignment. In the first version, *minimap*, which was simply designed to perform pairwise alignments, an alignment was approximated via a simple 1D-clustering of anchors to only determine whether to reads overlapped with each other [155]. Since *minimap2* also aims to perform base-level alignments, it first approximates an alignment by chaining anchors and penalizing the chains with would-be indels based on the size difference between pairs of anchors, similarly employed by the FASTP/FASTA method of Wilbur and Lipman [99]. As such, *minimap* and *minimap2* approximates both mapping of long-reads to a reference and pairwise alignment ,consequently speeding up both procedures since it avoids costly base-level alignment.

When base-level alignment is needed, the approximated segments can be subjected to a semi-global alignment via a two-piece affine gap penalty consisting of two different gap-opening and extension parameters corresponding for short and long indels—an adaptation of prior work from 1980-1990 [87–89, 95]. Specifically:

1

$$
\begin{aligned}
M_{(i,j)} &= max \begin{cases} M_{(i-1,j-1)} + score(a_i, b_j), \\ A_{(i-1,j)}, \\ B_{(i,j-1)}, \\ A'_{(i-1,j)}, \\ B'_{(i,j-1)} \end{cases} \\
A_{(i,j)} &= max\{M_{(i-1,j)} - o, A_{(i-1,j)}\} - e \\
B_{(i,j)} &= max\{M_{(i,j-1)} - o, B_{(i,j-1)}\} - e \\
A'_{(i,j)} &= max\{M_{(i-1,j)} - o', A'_{(i-1,j)}\} - e' \\
B'_{(i,j)} &= max\{M_{(i,j-1)} - o', B'_{(i,j-1)}\} - e'
\end{aligned}
\tag{1.5}
$$

Much like the affine-gap penalty allows one to identify large-indels by finding an optimal alignment path across three search spaces (e.g. $M_{(i,j)}, A_{(i,j)}, B_{(i,j)}$) the two-piece affine gap penalty has two additional search spaces, $A'_{(i+1,j)}$ and $B'_{(i,j+1)}$, utilizing different gap extension and gap opening penalties, e', and o', respectively. The additional search spaces allows an alignment to model a large structural event initially as a small insertion or deletion with additional mismatches (where the alignment score is maximum), but later the penalty becomes less costly due to the number of gaps leading to an alignment whose maximum score lies within the additional two search spaces. This enables *minimap2* to search for much larger indels without over-penalizing them, and indeed, enables one to search for intron and exon retention/skipping events in the alignment of RNA sequences. Furthermore, it also enables detection of smaller indels either due to true variation or sequencing errors. The boundaries of the initial approximated segments are also extend outwards using the Z-drop score, similar to BLAST's X-drop score.

Graphmap2

Graphmap2 [94], the successor to *Graphmap* [156], is also a long-read sequence mapper and aligner that fundamentally uses different approaches compared to *minimap2*. Both version rely on kmer matches, but are identified based on *gapped q-grams*, where each kmer is represented with multiple versions by altering pre-defined fixed positions. This technique is similar to that employed by Altschul et al. [101], aiming to increase the sensitivity of kmer matches by accounting for sequencing errors and biological variation. Similar to *minimap2*, the location of a sequence to a reference is then approximated by identifying kmer hits, but uses a 2D-space projection to a cartesian space to cluster and find collinear regions [156]. Anchors are then calculated via an edge-extended *de Bruijn* graph data structure and subjecting the approximated regions to a linear-walk in the graph. More specifically, a variant of the *de Bruijn* graph, $DB(s_i, k)$, is constructed for some chromosomal string in the reference, where for each vertex, s_i, a directed edge is added to d downstream vertices based on the starting position of s_i. Each vertex is also given a unique identifier, to prevent merging identical vertices preserving long-range context of s_i. As such, the anchors can be "mapped" to this graph data structure and an ordered walk through the graph can yield a noise-filtered set of anchors despite the high error profiles of long-reads (hence, the ability to traverse to up to d downstream kmers). Finally,

the resulting anchors between the read and the reference are curated by a variant of the longest-common sub-sequence algorithm which uses a flexible k-size string [156].

At this point, *Graphmap* and *Graphmap2* diverge: in both-cases, another round of anchor filtering is performed. In *Graphmap*, this is done via linear regression, while *Graphmap2* formulates it and solves it via the well-known 0-1 knapsack problem. With the resulting anchors, *Graphmap* performs base-level alignment is using some variant of a bit-vector-based alignment (introduced by Myers) with or without the affine-gap penalty. *Graphmap2* performs an affine or two-piece affine-gap penalty using the same library as that implemented by *minimap2*.

NGLMR

At the end of section 1.2.2, I mentioned the challenge of assembling non-haploid genomes. Although long-reads do provide an opportunity to generate haplotype-resolved *de novo* assemblers (which I discussed in more detail in section 1.1.4), inference of heterozygous structural variants is possible via careful partitioning of long-reads alignments in a reference.

The sequence mapper, *NGLMR*, was specifically designed to identify (heterozygous) structural variations [149]. Effectively, it breaks reads into smaller non-overlapping sub-sequences of a few hundred nucleotides (similar size to short reads), separately aligns them, and joins them together to identify positional breakpoints of structural variants. More specifically, the number of k-mer hits for a corresponding non-overlapping sub-sequence are computed via exact hashing, and candidate sub-sequences are only considered if the number of hits passes a threshold. These candidate sub-sequences undergo a base-level alignment and are considered (large) anchors if their similarity passes an additional threshold, which are then curated and chained by solving the longest increasing sub-sequence problem. At this point, the approach is similar BLAST's method to finding high-scoring segment pairs [101]. Additionally, *NGLMR* performs a final base-level alignment of anchors via a heuristic version of a Smith-Waterman alignment using concave gap-scoring penalty—similar to that of Waterman, Miller, and Myers [95, 96]. It also takes note of the distance between anchors in the read and the reference, since differences in their respective distances would indicate the presence of a structural variation.

MMseqs2

Third-generation sequencing technologies have further accelerated the growth of available whole-genome datasets of microbial organisms. As such, one particular topic in computational microbiology is *pan-genomes*—the variability and conservation of genes and gene-families in various microbial species. This often requires *de novo* identification of gene-families via a familiar problem: pairwise-alignments. Unlike long-read sequences, sequence variation between proteins are assumed to be only true evolutionary variation. And the definition of gene-families often employs a clustering-scheme based on respective sequence similarities. Thus, both fast and highly sensitive local and or global alignments are required.

MMseqs2 [154] utilizes a variety of algorithmic tricks to efficiently compute pairwise alignments and cluster billions of protein sequences in linear time. The *Linclust*-module of *MMseqs2* also uses a kmer-based anchoring, but unlike in DNA space where there can be $4k$ total kmers for some size of k, in protein space there are ($21k$ total kmers [157].

To reduce the search space, they reduce the number of amino acids to 13 by merging certain amino acids based on minimum mutual information of prior established likelihoods between amino acid substitutions (e.g. BLOSUM62 matrix). The search space if further reduced by only taking *m* smallest protein kmers as opposed to all kmers. Using these reduced profiles, *MMseqs2* effectively identifies sequences with matching kmer hits by sorting the kmer hash-table on kmer value. For each kmer group, the largest protein is chosen as the center sequence and it used to "recruit" other proteins to that center based on the position of the k-mer hits in their respective diagonals. These initial clusters are then curated by progressively more sensitive alignments including hamming distance, un-gapped alignment, and Smith-Waterman alignment. As such, the Linclust-module of MMseqs2 can cluster large protein sets by quickly identifying "draft clusters" and curating each cluster by more sensitive alignments within those clusters [154].

From software to hardware

A running theme in all these methods, is that direct and sensitive sequence alignment is computationally costly and needs to be avoided as much as possible if one wants to efficiently compare millions to billions of sequences. Aside from the various methods employed to reduce the number of required direct alignment operations, there has been a variety of both software and hardware algorithmic implementations for speeding-up sequence alignment. For example, all of the methods above use various optimization procedures such a banded global-alignment utilized by Wilbur and Lipman in 1985 [98], bit-vector implementation presented by Eugene Myers in 1999 [158], and some form of (improved) single-instruction, multiple data (SIMD) instructions to parallelize the dynamic programming procedure during the alignment [159]. There has also been efforts to offload sequence alignment to specialized hardware, such as FPGAs [160]. Alternatively, there are methods that completely avoid direct sequence alignment and only use approximations of them. These methods not only also scale to millions of (long-read) sequences, but also to millions of whole-genomes, further discussed in the next two sections, and are discussed in greater detailed in section 1.1.5.

1.1.4 Long-read *de novo* genome assembly

Third-generation sequencing entails that individual reads can span repetitive sequences in a genome yielding more complete assemblies [81, 109, 116]. This advantage has thus resurfaced interest in developing improved *de novo* assembly methods utilizing both overlap-layout-consensus and *de Bruijn* graph paradigms [109, 116]. Ultimately, the challenges imposed by third-generation sequencing are not entirely new: error rates, chimeric reads, and (large) repetitive sequences are still problematic. To some extent, the methods mentioned in the previous section facilitate handling of (high) sequencing errors, chimeric reads, and pairwise alignments bottle-neck. However, recent *de novo assemblers* employ new techniques to better construct the assembly graph (e.g. final string graph or *de Bruijn* graph), further discussed in this subsection.

The hierarchical genome assembly process (*HGAP*) in 2013 was among the first long-read only *de novo* genome assemblers, particularly designed for assembling sequencing data from *Pacific Bioscience* (PacBio) sequencers [122]. In short, *HGAP* followed the *Celera Genomics de novo* assembly scheme of error correction, assembly, and consensus. First,

HGAP took the longest reads up to some defined coverage and referred this subset as seed reads. All other reads were aligned to the seed reads to generate a consensus sequence via multiple-sequence alignment. This step not only corrected sequencing errors, but also discarded chimeric reads based on inconsistencies in coverage information. The corrected reads were then assembled using an existing overlap-layout-consensus assembler, such the *Celera Genomics* assembler, widely used in early genome sequencing projects such as *Drasophilia* [125] and humans [161]. Finally, remaining errors in the assembled contigs were "polished" by aligning all reads back to the assembly and generating a consensus sequences using a hidden Markov model from raw measurements (e.g. fluorescence and pulse information) stored during DNA sequencing via PacBio instruments.

Despite its clear advantages, third-generation sequencing in the first half of this decade was relatively more expensive than short-read sequencing (e.g. Illumina) [81, 109]. As such, a more economic strategy for *de novo* assembly was to sequence genomes in high-coverage with short-reads and complement the assembly with low-coverage long-read data [81, 109]. More specifically, *hybrid de novo* assemblers, such as *HybridSPades*, incorporated long-read information in an attempt to resolve complex subgraphs in the *de Bruijn* graph [162]. As previously discussed, a *de Bruijn* graph of short-reads (e.g. Illumina data) inherently harbours complex subgraphs due to repeat content and sequence heterozygosity, yielding less contiguous assemblies. However, by aligning long-reads to the *de Bruijn* graph via kmer seeding, the complex subgraphs could be disentangled by following only those supported by the long-reads, facilitating more accurate and contiguous assemblies. More specifically, *HybridSPades* first builds a *de Bruijn* graph from short-read data, and curates it to remove sequencing errors and artifacts, as well as low-supported nodes and edges [162]. Each long-read is then represented as a path in the *de Bruijn* graph based on valid path traversals of kmer matches [162]. Paths in the *de Bruijn* graph can therefore iteratively be traversed until no long-read consistently supports the traversal, resulting in a contig. Although this strategy could yield more contiguous assembly, it was still limited to the complexity of the *de Bruijn* graph, and thus repeat sequences and sequence heterozygosity were still challenging to resolve [162].

In the past few years, long-read-only *de novo* assembly became more affordable due to rapid developments in third-generation sequencing technologies, including increase in throughput and read-lengths, and lower error rates [81, 109]. Not surprisingly, many of the assemblers adapted to these improved third-generation sequencers use similar hierarchical strategies as *Celera Genomics* and *HGAP*, with a particular focus on reducing the computational bottleneck in pairwise sequence alignment for constructing an "error-free" overlap-graph.

Canu—an "updated" version of the *Celera Genomics* assembler—uses a variety of specialized pairwise sequence aligners along with internal heuristics (such as kmer frequency and read-coverage) to filter erroneous overlaps in the overlap-graph [151]. Canu uses *MHAP* [153] which avoids costly base-level sequence alignment by instead comparing kmer set-representations of reads using *MinHash* algorithm [163], discussed in more detail in section 1.1.5. In particular, it employs the *tf-idf* weight term frequency [164] to discriminate between repeat-induce and copy-number-induced overlaps, such as the case of microbes with high-copy number plasmids and/or non-haploid organisms [151]. *Canu* iteratively corrects reads via consensus and trimming and uses these reads to identify valid

overlaps within a defined number of standard deviations of the global-error rate (e.g. the sequence divergence of two overlapping reads is higher the median sequence divergence of all overlaps) and coverage of a read (e.g. non-chimeric reads). The remaining overlaps and corresponding reads are then used to construct a *best overlap graph*, which is a greedy approach to the overlap graph using only the longest overlaps for each read as a means to reduce memory requirements [151, 165]. It then generates a candidate set of contigs by identifying non-conflicting paths in the best overlap graph by greedily ranking reads based on "reachability" (e.g. number of other reads reachable to it) and using the rankings to guide the traversals [151, 165]. Although "bubbles" may arise due to sequence and structural heterozygosity in the corresponding genome, only the longest path is retained [151, 165]. The resulting contigs are finally curated by "breaking" sequence if there are evidence of unresolved repeats or branching.

Despite its robustness and accuracy, *Canu* suffers from high runtime and computational tractability, such as > 100 GB of memory and several weeks of wall-clock even after parallelization when assembling a human genome [147, 152, 166][. This is largely due to its precise read and kmer indexing as well as iterative error correction modules [147, 151]. Alternatively, *miniasm*—a companion to the long-read sequence aligner, *minimap* and *minimap2*, discussed in the previous section—reduces the runtime and computational requirements by avoiding error-correction [155]. It uses an optimized pairwise sequence aligned to generate the overlap graph (*minimap* and *minimap2* as described in section 1.1.3) to quickly identify non-chimeric overlaps with sufficient overlap length and constructs a variant of the string graph where the vertex set are reads [155]. The difference here is that base-level accuracy of the assembly reflects that to the error-rate of the input sequencing data, as reads are not error-corrected (except for trimming), and contigs are generated by representing only one path after collapsing "small-enough" bubbles. Despite of this, *miniasm* still achieves contiguous assemblies compared to *Canu* with fewer resources and shorter run-times [155, 167]. As such, methods such as *RACON* and *Pilon* have been used to provide high base-level accuracy from *miniasm* assemblies [91, 167, 168].

Aside from an overlap-layout-consensus approach, there have also been methods utilizing *de Bruijn* graphs long-read *de novo* assembly [121, 169]. Particularly, the *ABruijn* assembler is an adapted version of the *HybridSPades* assembler. Unlike a *de Bruijn* graph, the edges set for ABruijn graph describe the positional offset of the starting locations of two nodes [121]. Thus, two k-1 nodes can be connected despite not being adjacent with each other in a string. An *ABruijn* graph can therefore be constructed from the long-read data set by choosing a k-mer length. k, and frequency threshold, t, that maximize the number of unique kmers [121]. Similar to *HaybridSPades*, by representing each long-read as a path in the *ABruijn* graph, contigs can be generated through an iterative traversal until the long-reads no longer consistently support it [121]. A similar approach can then be used to correct errors and generate consensus sequence for the resulting contigs [121]. The advantage of the *ABruijn* graph is that it does not require pairwise read-alignments of the long-read dataset, significantly speeding up the time and computational resources needed to construct an assembly [121, 147, 152, 166]. However, resolution of repeat sequences are still limited to read-length [121], and the sensitivity of the assembly as a whole is heavily depends consistent unique kmers throughout the long-read dataset [121], which can be negatively influenced by the error prone nature of current third generation sequencing

technologies [81, 109].

As previously discussed, organisms with non-haploid genomes can harbour variation within the same chromosome copy. As such, the genomes of these organisms have been historically challenging to assemble as most assemblers generate a haploid representation from the string-graph [137, 139–141]. Analogous to repetitive sequences, heterozygous variants in the graph—encoded as "bubbles"—can in theory be disentangled and assign to one of the chromosome copies, given that there are reads long-enough to bridge across the bubbles and anchor to unambiguous sequences. In fact, with sufficient length, long reads can span across stretches of variation and traverse through unique paths across the bubbles.

FALCON-UNZIP (in conjunction with *FALCON* assembler) aims to exploit the topological features of heterozygous structural variants in the graph to provide *de novo* haplotype resolution [170]. As with all overlap-layout-consensus assemblers, both use a hierarchical approach of error-correction, contig assembly, and consensus sequence generation [122, 125, 151, 165]. However, careful heuristics are used to identify and separate haplotypes. When constructing the string graph, *FALCON-UNZIP* tries to identify signals of heterozygous events, particularly *simple paths* (a subgraph starting at a unique source and sink node with no internal branching—and *compound paths* (a set of overlapping simple paths, but also with a unique source and sink node). By virtually "linearizing" compound and simple paths, contigs can be generated through a non-conflicting path finder, as previously mentioned [122, 125, 151, 155, 165], but the contigs retain their heterozygous information. For each compound or simple path in a contig, *FALCON-UNZIP* returns to the original overlaps used to construct the string graph and partitions them to *block-phase identifiers*, representing groups of reads supporting some specific path in the assumed heterozygous-induced bubbles. It then reconstructs a string graph for that contig ignoring overlaps of reads with conflicting block-phase identifiers and identifies linear paths which correspond to different haplotype-resolved sequences. Despite these improvements, haplotype-resolved genomes for organisms with more than two chromosome copies remain challenging to resolve beyond diploid genome configuration [135, 138, 149, 170–172].

Nevertheless, third-generation sequencing technologies are enabling assemblies of higher quality than those obtained via NGS data alone. Although third-generation sequencing at the moment is relatively more expensive than NGS, one can imagine complementing the architectural information of a one more long-read genome assemblies with population information of hundreds to thousands of individuals using (existing) short-read sequencing data. Ultimately, the price of third-generation sequencing should drop in the next coming years to a level comparable to NGS. This will enable large collections of high-quality assemblies, opening novel opportunities to compare whole genomes together in their entirety, as opposed to relying on a single reference genome. Indeed, these ideas are currently being discussed in the bioinformatics community [173, 174].

Ultimately, third-generation sequencing adds to the ever increasing mountain of available genomic data. As such, a variety of "approximate" methods have been developed for comparing and mining genomic information in both long and short-read datasets. These methods optimize for speed at the cost of optimal solutions, as they primarily aim to scale from hundreds to thousands of genomes. In the next section, I provide an overview of these "approximate" or *genomic-streaming* methods.

1

1.1.5 Genomic fingerprints

The first instance of so-called "approximate" methods can be seen a few years before the introduction dynamic programming for sequence alignment. Between the 1960-1970, efforts to establish computational methods to compare sequences emphasized the importance of statistical statements on whether two sequences were significantly related (e.g. homology). Walter Fitch in 1966 devised a method to determine the similarity of two protein sequences and its statistically significance based on pairwise-distances between every possible k-mer in each sequence [175]. More specifically, the distance between two k-mers was defined as the number of single nucleotide DNA changes needed to transform one k-mer into the other (keep in mind that this was only made possible due breakthroughs in unlocking the genetic code between DNA codons and amino acids). Based on a null distribution that would be generated from the pairwise distances of randomly generated sequences of similar length, Fitch could test whether the observed and null distributions were indeed significantly differently [175]. In other words, Fitch's method derived both similarity score and homology significance based purely on k-mer content. Fifty years later, computing similarities using only k-mer content-information has become particularly popular when analysing large datasets of long-read sequences and/or genomes due to scalability promises.

MHAP, for example, is an aligner designed to perform the critical pairwise alignment step in *de novo* assembly [153]. Rather than employing direct sequence alignment, it instead uses the *MinHash* algorithm, first proposed as an approximate but fast way to compare documents and websites in the world-wide web [163]. This algorithm has been the basis for a variety of genomic-streaming models, so I provide a quick overview of it:

The idea of *MinHash* is to not compare documents in their entirety, but to compare only their "fingerprints". More specifically, let two sets, $Set(A)$ and $Set(B)$ represent the set of words found in each document, A and B, respectively. The overall similarity of the two documents can approximated via the Jaccard Index, $JI(A,B) = \frac{|Set(A) \cap Set(B)|}{|Set(A) \cup Set(B)|}$ [163]. In other words, comparing the fraction of shared words over total words from the two set can approximate the overall similarity of A and B. Logically, documents can have thousands of words, and so the memory and time requirements are linearly dependent on the document sizes. This may not be a problem if you are only comparing a small number of documents, but in the world-wide web, with million-to-billions of websites, a naïve-implantation will not scale. Instead, Andrei Broder proposed the idea of *MinHash*: hashing every word with n hash-functions, $H_1, H_2, \cdots, H_n$, and computing the Jaccard Index as the number of instances where the smallest hash for each H_i of the two documents are equal, greatly reducing the memory and time requirements since generally $n << |Set(A)|, |Set(B)|$. Alternatively (and more relevant in genomics), one can use a single hash-function and obtain the n-smallest hashes between two documents—referred to as the *bottom-sketch*. These bottom-sketches act as "fingerprints" of the original documents, whose size are significantly smaller than the size of the original set of words. However, the similarity of the documents can still be approximated via a Jaccard-Index of their sketch-representations, where the accuracy is bounded by the size of sketch [163].

It's not hard to see how *MinHash* can be applied to genomics: sequences and genomes (e.g. documents) can be represented by their kmer content (e.g. words). Recall that $H'(k_i) = H(smallest\{k_i, \overline{k_i}\})$, represent the hash-value of the lexicographic-smallest se-

quence between kmer k_i and it s reverse-complement $\overline{k_i}$, or the smallest hash-value of the two. The bottom-sketch for a set of sequences, S, a kmer size, k, and a sketch-size, l, can be constructed by taking the l smallest hash values from the set of all k-mers in a genome, denoted as $BottomSketch(S, k, l)$.

Indeed, *MHAP* exploits this property by representing each long-read with its bottom-sketch and approximating their sequence similarity via the corresponding Jaccard Index [163], avoiding costly base-level sequence alignment.

Around the same time, Huan Fan *et al.* in 2015 proposed a statistical foundation to make the Jaccard Index of kmer sets more biologically relevant [176]. The main aim of the method was to obtain biologically-relevant distances between genomes that can be used for phylogenetic reconstruction, without needing to perform (whole-genome) sequence alignments. Aside from the high computational cost of whole-genome alignment, *de novo* assemblies are inherently an incomplete representation for non-haploid genomes, as previously discussed in section 1.1.2 and 1.1.4. Therefore the original sequencing dataset (e.g. R) can be more informative than its corresponding assembly. Similar to *MHAP*, the authors compute the (exact) Jaccard-Index of two kmer sets, although their interpretation differed. If $Set(G_1)$ and $Set(G_2)$ represent the kmer sets for genomes G_1 and G_2, respectively, then $JI(Set(G_1), Set(G_2)) \approx e^{-k\epsilon}$ [176]; the right-hand of the equation being the probability that no kmer undergoes a mutation under a Poisson-distribution, where ϵ here serves as the average sequence divergence of the two genomes (analagous to the original definition of the error-rate of two sequences). Therefore, the average sequence divergence between genomes can be estimated, $\epsilon \approx \frac{-1}{k} JI(Set(G_1), Set(G_2))$ [176]. Importantly, this assumes that genome size of G_1 and G_2 are equal, in other words, do not account for indels events which decrease/increase their respective kmer-set sizes. Therefore, Fan *et al.* addressed indels by replacing $|Set(G_1) \cup Set(G_2)|$ with the largest of the two. Ambitiously, Fan *et al.* computed the exact Jaccard-Index of all k-mers observed in the collection of sequencing datasets being analyzed, making it computationally expensive to scale for more than a dozen genomes [176].

A more scalable version of the method proposed by Fan *et al.* was adopted by Ondov *et al.* in 2016 [177]. Instead of representing a genome or read-set with all corresponding kmers, their method, *MASH*, uses their corresponding bottom-sketches to estimate ϵ, while basing the Jaccard-Index on the average size of two genomes, thus leading to a slightly different approximation, $\epsilon \approx \frac{-1}{k} ln \frac{2j}{1+j}$, where j is the Jaccard-Index using the bottom-sketches of G_1 and G_2. Since then, there has been a variety of different sketching algorithms and applications utilizing different algorithmic techniques such as count-min sketches for k-mer frequency estimations, hyper-log log for the number of distinct number kmers in a genome, adaptations of bitsets for fast sequence searching [178], all which are further discussed in a recent review [107].

Despite the recent development of sketching algorithms in genomics, the idea of genomic "fingerprints" draws parallels to landmark experimental studies dating back many decades before.

In 1977, Carl Woese and George Fox used a method embodying the same principles behind genome comparison via k-mer sets [179]. Notably, this method was used to provide the first evidence of a separate domain of living organisms: archaea. At the time, it was

1

generally accepted that two eukaryotes and bacteria were the two kingdoms of life, largely driven by observed differences in their cell organization such as the presence/absence of a nucleus [179]. Although some bacterial specimens were phenotypically different from bacteria, such as methanogens containing unique enzymes that produce methane, they were generally regarded as a sub-group of bacteria [179].

Woese and Fox in 1977 argued that one cannot simply compare organisms by the morphology/organization of a cell, and since evolutionary history recorded itself in the genomes of organisms, true evolutionary relationship is better inferred by analyzing genomes [179]. At the time, DNA sequencing technology was still in its early stage, as Frederick Sanger and Alan Coulson were beginning to pioneer early version of first generation sequencing technology in 1975 [180]. So rather than (impossibly) sequencing the entire genome of various microbial organisms, Woese and Fox instead focused on sequencing the 16s ribosomal gene in a set of bacteria, eukaryotes and methanogens specimens, since the 16s gene was observed to be present in all bacteria and eukaryotes (in eukaryotes, the equivalent is 18s). Importantly, they did not aim to reconstruct the entire 16s gene, but instead generate "fingerprints" from each specimen such that they can then be compared.

More specifically, Woese and Fox isolated 16s (18s) RNA from the different organisms and subjected them to an T1 RNase, which digested RNA molecules at every guanosine nucleotide, effectively generating kmers of various sizes. The resulting sub-sequences were then sequenced, represented each organism as the set of sub-sequences from the RNase digestion, and clustered them using a similarity score resembling the Jaccard Index, $S(a,b) = \frac{2N_{AB}}{N_A+N_B}$ [179], where N_A and N_B are the number of sequences in organisms A and B, respectively, and N_{AB} the number of sequences shared between A and B. The experiment showed that not only did bacteria and eukaryotes clustered separately, but that the methanogens formed their own cluster equally distant to eukaryotes and bacteria. As such, Woese and Fox proposed that methanogens form their own domain, which they regarded as textitarchae [179]; a view that is ultimately supported today.

An additional example of the application of early genomic "fingerprints" can be found nearly three decades before Woese and Fox. In 1950, Pehr Edman developed the first sequencing technique designed for protein sequencing [181]. Famously known as *Edman degradation*, a given protein was chemically "degraded" into smaller fragments of a few amino acids using phenyl isothiocyanate, allowing one to "cleave" proteins by lowering the PH [181]. The fragments can then be individually analyzed via paper chromatography to determine their amino acid order and composition [181, 182]. The first set proteins were thus sequenced in the 1950s, including multiple protein-chains in insulin [182–184] and "healthy" and sickle-cell versions of human haemoglobin [185, 186]. It also during this decade when James Watson and Francis Crick in 1953 published their famous paper on the double-helix nature of DNA, providing an explanation to how DNA can store and pass genetic information across generations [187]. Shortly after, Francis Crick's famous lecture and publication, *On protein synthesis*, in 1957 and 1958, respectively, postulated that DNA was responsible for the synthesis of proteins, which he believed was mediated by an unknown molecular mechanism (which we now as translation) [75, 188]. He postulated the *central dogma of molecular biology* on how information, irreversibly, flowed from DNA to proteins, and foresaw that by comparing DNA and protein sequences, one could unravel both the evolutionary history and molecular basis for the phenotype of an organism [75,

188].

As such, ideas behind the relevance genomic sequence comparison were beginning to take form. This was particularly re-enforced by Vernon Ingram in 1956-1957, when he provided evidence that variation in the sequence composition of the haemoglobin protein was linked to sickle-cell disease, illustrating one of the first concrete connections between molecular biology and human disease [185, 186, 189]. At the time, Ingram was not able to directly sequence haemoglobin and analyse its sequence composition, and thus, could not compare the sequence composition of the protein. Instead, in his 1956 study, he cleverly compared the proteins by constructing what he called an experimental "fingerprint" of the normal and sickle-cell-versions haemoglobin [185]. He did this by separately breaking down the two proteins into fragments (which, similar to Woese and Fox, serve as k-mers of various sizes), and running them on a two-dimensional electrophoresis gel, allowing separations of the fragments based on the charged-properties of their amino acid composition [185]. The resulting image (which can be obtained via paper chromatography) served as a fingerprint of the proteins [185]. For normal and sickle-cell haemoglobin, he found that their fingerprints were nearly identical with the exception of one fragment [185]. Encouraged by these results, Ingram sequence the both versions haemoglobin proteins in the following year and showed that that they differed by a single position, where a glutamic acid was replaced by a valine [186], and thus illustrating how simple changes in a genome can give rise to human diseases [186].

1.1.6 Microbial pan-genomes

As noted multiple times throughout the past few sections, long-read sequencing data provides a feasible way to obtain complete microbial genomes. As such, investigations of microbial *pan-genomes*—that is, the collective genomic sequences, gene-content, and structural organization in a microbial population—become particularly interesting, since microbes routinely shed an integrate new genomic sequences. On one end of the spectrum, there are microbes with little pan-genome variation due to the nature of their isolated habitats [190, 191]. On the extreme end, two strains from the "species" can share as little as 10% of their genomic content, resulting in large variations in their pan-genomes [192]; ultimately highlighting (philosophical) issues on the discretisation and definition of microbial species. Nevertheless, comparing the genomes of multiple microbes is naturally a computational task, and the methods behind them are often developed in context of some specific biological and/or evolutionary question.

One such example is *FAST-ANI* [193], which employs the *MASH-map* method under-the-hood. The main goal of this method is to quickly calculate average-nucleotide identities (ANI) across large collections of microbial genomes—a metric often used to assess evolutionary distances and species boundaries based on whole-genome data (e.g. *de novo* assemblies). Computationally, it requires identification of homologous regions/genes between genomes and calculating their average similarity, which is ultimately a two-step pairwise whole-genome/sequence alignment. Traditionally, the calculations are based on direct sequence alignment, which as already discussed, is computationally costly and demanding to scale. As such *FAST-ANI* approximates the ANI metric between two genomes by fragmenting one to non-overlapping sub-sequences of a few thousand base-pairs (analogous to generating long-read-versions of a genome) and mapping them using *MASH-*

map [193]. A general ANI score between two genomes can thus be calculated by heuristically binning the alignments onto a genome and averaging their sequence similarities. Although an approximation, it enable comparison of more than 90,000 bacterial genomes, and with this large sample size, it suggesting a species boundary between genomes with an ANI of 83-95% [193].

On a more specific application, several methods have aimed to automate the construction of a pan-genomes by identifying all gene-families in a collection of genomes and noting their conservation and variation. *Roary* [194], *BPGA* [195], *Panseq* [196], *PanGP* [197], *PanOCT* [198] , *PGAP* [199], and *ITEP* [200] are all computational frameworks that aims to do just that: they perform pairwise-gene alignments using some version of their favourite sequence aligner, and (iteratively) cluster the genes based on sequence similarity to identify conserved and variable gene families. In principle they aim to facilitate more concrete down-stream analysis, such as evolutionary inference on gene-content, gene-family enrichment, and gene-family exclusivity.

An example of a concrete down-stream analysis based on a defined pan-genome is *Scoary*, which aims to perform genome-wide associations (GWAS) on the presence/absence of gene families [201]. Potentially, GWAS can reveal how the acquisition of one or more new gene-families in a population leads to a new phenotype, such as acquired drug-resistance. As such, *Scoary* processes the gene-families of a population of genomes and performs a first-pass Fisher's exact test to identify candidate genes that are significantly associated with sub-groups of genomes that share a common phenotype. These are then further analyzed by accounting for the population structure based on a phylogenetic tree and computing a test-statistic based on the distribution of gene-families across the tree. However, these analyses are only based on presence/absence of "similar" genes, and do not discern potential sequence variations that may drive alternate protein functions.

1.2 An overview of this thesis

I started this introductions by discussing how yeast (and it alcoholic by-products) influenced human evolution, modern societies, and the biological sciences. I then transitioned into a semi-historical overview of bioinformatics algorithms, focusing on sequence analysis, *de novo* assembly, and comparative genomics in the era of long-read sequencing. So how do the these topics relate to the academic work of this thesis?

1.2.1 The case of the missing MAL gene

Despite more than a century worth of scientific research, we are still actively understanding the global genomic diversity *Saccharomyces* species. But this has not stopped efforts in genetically engineering yeast for industrial applications. One particular industrial yeast is *CEN.PK113-7D*: a strain that can thrive in industrial conditions, originating from various genetic crossings of other *S. cerevisiae* strains in the 1990s [141, 202–205]. Despite only having a haploid genome, it wasn't until the 2010-2012, when the first *de novo* assemblies were produced [141, 206]. Aside from issues regarding the fragmented and incomplete quality of the assemblies, Nijkamp *et al.* pointed out one particular inconsistency in the results: experimentally, the industrial-relevant gene, the MAL locus (which enables yeast to process specific sugars), highlighted four distinct copies in the genome of *CEN.PK113-7D* [141]. However, computational results showed only three. Nijkamp *et al.* argued the inconsistency in the copy-number of this loci was due to a collapse repeat (see section 1.1.2), and that the additional copy indeed existed somewhere in the genome of *CEN.PK113-7D*.

In chapter 2, I sequenced and assembled the genome of *CEN.PK113-7D* using novel long-read sequencing technology, unraveling an unexpected lesson in microbial genome evolution.

1.2.2 Tracing genome mosaicism in microbial genomes

As alluded in the previous section, *Saccharomyces* species often undergo natural crossings and hybridization events with other yeasts, ultimately increasing their genetic diversity and phenotypic attributes. As such, strain engineering efforts often employ a form to artificial crossing and hybridization with other (selected) yeast strains, as was the case for *CEN.PK113-7D*. Regardless on whether the events are natural or artificial, both ultimately lead to integration of new alleles and sequences, resulting in *mosaic genomes*, that is, genomes with multiple evolutionary origins.

At the end of chapter 2, we performed a simple computational analysis for identifying the global origins of various subsequences *CEN.PK113-7D*, ignoring contextual information throughout a chromosome [207].

In chapter 3, I developed a kmer-based method aiming to systematically trace the origins of a (mosaic) genome across all chromosomes, guided by (large) collections of available sequencing datasets of *Saccharomyces* genomes.

1.2.3 Where do lager-yeast originate?

Natural and artificial crossings make *Saccharomyces* species particularly challenging to study from an evolutionary perspective. One example is *S. pastorianus*: a species derived from a natural crossing of *S. cerevisiae* and *S. eubayanus*, which gave rise to lager-beer

1

brewing. Despite its overwhelming popularity in the alcoholic world, the exact origins of *S. pastorianus* are unclear, but it boils down to two proposed hypothesis: (1) the species arose from a single hybridization event between *S. cerevisiae* and *S. eubayanus*, or (2), the species arose from multiple hybridization events between *S. cerevisiae* and *S. eubayanus* [208]. To make this question even more challenging, the genomes of *S. pastorianus* are aneuploid: each chromosome has a non-uniform distribution fo chromosome copies, ranging from 0 to more than 5 [137, 209]. Various studies have employed whole-genome sequencing to better understand their genomic landscape and evolutionary history, but the aneuploid nature of their genomes leads to highly fragmented assemblies complicating down stream analysis [137, 208, 210, 211].

In chapter 4, I sequenced a strain of *S. pastorianus* using long-read sequencing technology, allowing us to obtain the most complete *de novo* assembly of this species [212] (at least when it was first published). Using the method developed in chapter 4, we computationally tested the competing hypothesis of their evolutionary origins.

1.2.4 A streaming algorithm to infer species-composition in *Saccharomyces* genomes

Rapid improvements in whole-genome sequencing technologies is enabling scientists to discover (natural) *Saccharomyces* hybrid-genomes. It is therefore not difficult to foresee screening campaigns by industrial and academic institutions who aim to unravel the global genomic diversity of these yeasts.

In chapter 5, I developed an alignment-free streaming algorithm to infer hybrid-species composition in *Saccharomyces* genomes. The algorithm quickly identifies the presence of one or more species from the *Saccharomyces sensu strictu* and approximates their relative genomic contribution, facilitating downstream genomic characterizations and evolutionary analysis.

1.2.5 How can one compare *n* diverse microbial genome assemblies?

Throughout this thesis, I frequently discuss how long-read sequencing enables more complete reconstructions of microbial genomes. The previous chapters focused on downstream analysis in a single *de novo* assembly. But what if one wants to analyse multiple microbial genomes?

In chapter 6, I focused in the computational task of comparing genome architectures—that is, the order and arrangement of genes in a genome—across different microbial populations. Long-read assemblies enable us to identify differences and similarities in local and global operon structures, as well as high-level structural variation. As such, I developed a method that represents a collection of genomes as a gene-based multi-directed graph, enabling simultaneous comparison of microbial proteomes with little to extreme genomic diversity [213].

1.2.6 Can we better educate microbiologists in bioinformatics?

Finally, the introduction of this thesis focused on selected origins of microbiology (e.g. yeasts and alcohol) and bioinformatics (e.g. sequence analysis and comparative genomics). Together, these two topics can serve as a powerful starting point to train the next generation of scientists.

In this chapter, I implemented a novel curriculum to educate Bachelor-level microbiologists with cutting-edge advancements in bioinformatics [214]. As is the running theme of this thesis, the curriculum focused on long-reads and comparative genomics. Specifically, students used their existing laboratory techniques to sequence novel bacterial organisms using long-read sequencing technologies. Through the data that they personally generate, we interactively taught fundamental topics of algorithms in bioinformatics, and how they can be utilized to better understand microbiology.

2

Nanopore sequencing enables near-complete de novo assembly of *Saccharomyces cerevisiae* reference strain CEN.PK113-7D

The haploid Saccharomyces cerevisiae strain CEN.PK113-7D is a popular model system for metabolic engineering and systems biology research. Current genome assemblies are based on short-read sequencing data scaffolded based on homology to strain S288C. However, these assemblies contain large sequence gaps, particularly in subtelomeric regions, and the assumption of perfect homology to S288C for scaffolding introduces bias. In this study, we obtained a near-complete genome assembly of CEN.PK113-7D using only Oxford Nanopore Technology's MinION sequencing platform. Fifteen of the 16 chromosomes, the mitochondrial genome and the 2-µm plasmid are assembled in single contigs and all but one chromosome starts or ends in a telomere repeat. This improved genome assembly contains 770 Kbp of added sequence containing 248 gene annotations in comparison to the previous assembly of CEN.PK113-7D. Many of these genes encode functions determining fitness in specific growth conditions and are therefore highly relevant for various industrial applications. Furthermore, we discovered a translocation between chromosomes III and VIII that caused misidentification of a MAL locus in the previous CEN.PK113-7D assembly. This study demonstrates the power of long-read sequencing by providing a high-quality reference assembly and annotation of CEN.PK113-7D and places a caveat on assumed genome stability of microorganisms.

2.1 Introduction

Whole genome sequencing (WGS) reveals important genetic information of an organism that can be linked to specific phenotypes and enable genetic engineering approaches [215] [216]. Short-read sequencing has become the standard method for WGS in the past years

This chapter has been published in FEMS Yeast Research [207]

due to its low cost, high-sequencing accuracy and high output of sequence reads. In most cases, the obtained read data is used to reassemble the sequenced genome either by de novo assembly or by mapping the reads to a previously assembled closely related genome. However, the sequence reads obtained are relatively short: between 35 and 1000 bp [217]. This poses challenges as genomes have long stretches of repetitive sequences of several thousand nucleotides in length and can only be characterised if a read spans the repetitive region and has a unique fit to the flanking ends [218]. As a result, de novo genome assembly based on short-read technologies 'break' at repetitive regions preventing reconstruction of whole chromosomes. The resulting assembly consists of dozens to hundreds of sequence fragments, commonly referred to as contigs. These contigs are then either analysed independently or ordered and joined together adjacently based on their alignment to a closely related reference genome. However, reference-based joining of contigs into so-called scaffolds is based on the assumption that the genetic structure of the sequenced strain is identical to that of the reference genome—potentially concealing existing genetic variation.

Previous genome assemblies of the *Saccharomyces cerevisiae* strain CEN.PK113-7D have been based on homology with the fully assembled reference genome of *S. cerevisiae* strain S288C [141, 219]. CEN.PK113-7D is a haploid strain used as a model organism in biotechnology-related research and systems biology because of its convenient growth characteristics, its robustness under industrially relevant conditions and its excellent genetic accessibility [141, 202, 220, 221]. CEN.PK113-7D was sequenced using a combination of 454 and Illumina short-read libraries, and a draft genome was assembled consisting of over 700 contigs [141]. After scaffolding using MAIA [222] and linking based on homology with the genome of S288C, it was possible to reconstruct all 16 chromosomes. However, there were large sequence gaps within chromosomes and the subtelomeric regions were left unassembled, both of which could contain relevant open reading frames (ORFs) [141]. Assuming homology to S288C, more than 90% of missing sequence was located in repetitive regions corresponding mostly to subtelomeric regions and Ty-elements. These regions are genetically unstable as repeated sequences promote recombination events [223]; therefore, the assumption of homology with S288C could be unjustified. Ty-elements are present across the genome: repetitive sequences with varying length (on average ~6 Kbp) resulting from introgressions of viral DNA [224]. Subtelomeric regions are segments towards the end of chromosomes consisting of highly repetitive elements making them notoriously challenging to reconstruct using only short-read sequencing data [225]. While Ty-elements are likely to have limited impact on gene expression, subtelomeric regions harbour various so-called subtelomeric genes. Several gene families are present mostly in subtelomeric regions and typically have functions determining the cell's interaction with its environment, such as nutrient uptake [226–228], sugar utilisation [229] and inhibitor tolerance [230]. Many of these subtelomeric gene families therefore contribute to the adaptation of industrial strains to the specific environment they are used in. For example, the RTM and SUC gene families are relevant for bioethanol production as they increase inhibitor tolerance in molasses and utilisation of extracellular sucrose, respectively [226, 230]. Similarly, MAL genes enable utilisation of maltose and maltotriose and FLO genes enable calcium-dependent flocculation, both of which are crucial for the beer brewing industry [231–233](Teunissen and Steensma 1995; Lodolo et al.2008; Brown, Murray

and Verstrepen 2010). As is the case for Ty-elements, subtelomeric regions are unstable due to repetitive sequences and homology to various regions of the genome, which is likely to cause diversity across strains [141, 223, 233]. Characterising and accurately localising subtelomeric gene families is thus crucial for associating strain performance to specific genomic features and for targeted engineering approaches for strain improvement [225].

In contrast to short-read technologies, single-molecule sequencing technologies can output sequence reads of several thousand nucleotides in length. Recent developments of long-read sequencing technologies have decreased the cost and increased the accuracy and output, yielding near-complete assemblies of diverse yeast strains [234, 235]. For example, de novo assembly of a biofuel production *S. cerevisiae* strain using PacBio reads produced a genome assembly consisting of 25 chromosomal contigs scaffolded into 16 chromosomes. This assembly revealed 92 new genes relative to S288C amongst which 28 previously uncharacterised and unnamed genes. Interestingly, many of these genes had functions linked to stress tolerance and carbon metabolism that are functions critical to the strains industrial application [234]. In addition, rapid technological advances in nanopore sequencing have matured as a competitive long-read sequencing technology and the first yeast genomes assembled using nanopore reads are appearing [234–238]. For example, Istace *et al.* sequenced 21 wild *S. cerevisiae* isolates and their genome assemblies ranged between 18 and 105 contigs enabling the detection of 29 translocations and four inversions relative to the chromosome structure of reference S288C. In addition, large variations were found in several difficult to sequence subtelomeric genes such as CUP1, which was correlated to large differences in copper tolerance [237]. Nanopore sequencing has thus proven to be a potent technology for characterising yeast.

In this study, we sequenced CEN.PK113-7D using Oxford Nanopore Technology's (ONT) MinION sequencing platform. This nanopore de novo assembly was compared to the previous short-read assembly of CEN.PK113-7D [141] with particular attention for previously, poorly assembled subtelomeric regions and for structural variation potentially concealed due to the assumption of homology to S288C.

2.2 Materials and Methods

2.2.1 Yeast strains

The *Saccharomyces cerevisiae* strain 'CEN.PK113-7D Frankfurt' (MATa MAL2-8c) was kindly provided by Dr P. Kötter in 2016 [141, 205]. It was plated on solid YPD (containing 10 g/l yeast extract, 20 g/l peptone and 20 g/l glucose) upon arrival, and a single colony was grown once until stationary phase in liquid YPD medium and 1 mL aliquots with 30% glycerol were stored at -80℃ since. The previously sequenced CEN.PK113-7D sample was renamed 'CEN.PK113-7D Delft' [141]. It was obtained from the same source in 2001 and 1 mL aliquots with 30% glycerol were stored at -80℃ with minimal propagation since (no more than three cultures on YPD as described above).

2.2.2 Yeast cultivation and genomic DNA extraction

Yeast cultures were incubated in 500-mL shake flasks containing 100 mL liquid YPD medium at 30°on an orbital shaker set at 200 rpm until the strains reached stationary phase with an OD660 between 12 and 20. Genomic DNA of CEN.PK113-7D Delft and CEN.PK113-7D

Frankfurt for WGS was isolated using the Qiagen 100/G kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions and quantified using a Qubit®Fluorometer 2.0 (ThermoFisher Scientific, Waltham, MA, USA).

2.2.3 Short-read Illumina sequencing

Genomic DNA of CEN.PK113-7D Frankfurt was sequenced on a HiSeq2500 sequencer (Illumina, San Diego, CA) with 150 bp paired-end reads using PCR-free library preparation by Novogene Bioinformatics Technology Co., Ltd (Yuen Long, Hong Kong). All Illumina sequencing data are available at NCBI (`https://www.ncbi.nlm.nih.gov/`) under the bioproject accession number PRJNA393501

2.2.4 MinION sequencing

MinION genomic libraries were prepared using either nanopore Sequencing Kit SQK-MAP006 (2D-ligation for R7.3 chemistry), SQK-RAD001 (Rapid library prep kit for R9 chemistry), or SQK-MAP007 (2D-ligation for R9 chemistries) (Oxford Nanopore Technologies, Oxford, UK). Two separate libraries of SQK-MAP006 and one library of SQK-RAD001 were used to sequence CEN.PK113-7D Delft. Only one SQK-MAP007 library was used to sequence CEN.PK113-7D Frankfurt. With the exception of the SQK-RAD001 library, all libraries used 2-3 μg of genomic DNA fragmented in a Covaris g-tube (Covaris) with the '8—10 kbp fragments' settings according to manufacturer's instructions. The SQK-RAD001 library used 200 ng of unsheared genomic DNA. Libraries for SQK-MAP006 and SQK-MAP007 were constructed following the manufacturer's instructions with the exception of using 0.4× concentration of AMPure XP Beads (Beckman Coulter Inc., Brea, CA, USA) and 80% EtOH during the 'End Repair/dA-tailing module' step. The SQK-RAD001 library was constructed following the manufacturer's instructions. Prior to sequencing, flow cell quality was assessed by running the MinKNOW platform QC (Oxford Nanopore Technology). All flow cells were primed with priming buffer and the libraries were loaded following the manufacturer's instructions. The mixture was then loaded into the flow cells for sequencing. The SQK-MAP006 library of CEN.PK113-7D Delft was sequenced twice on a R7.3 chemistry flow cell (FLO-MIN103), and the SQK-RAD001 library was sequenced on a R9 chemistry flow cell (FLO-MIN105)—all for 48 h. The SQK-MAP007 library for CEN.PK113-7D Frankfurt was sequenced for 48 h on a R9 chemistry flow cell (FLO-MIN104). Reads from all sequencing runs were uploaded and base-called using Metrichor desktop agent (`https://metrichor.com/s/`). The error rate of nanopore reads in the CEN.PK113-7D Frankfurt and Delft was determined by aligning them to the final CEN.PK113-7D assembly (see section below) using Graphmap (Sović et al.2016) and calculating mismatches based on the CIGAR strings of reads with a mapping quality of at least 1 and no more than 500 nt of soft/hard clipping on each end of the alignment to avoid erroneous read alignments due to repetitive regions (i.e. paralogous genes, genes with copy number variation). All nanopore sequencing data are available at NCBI under the bioproject accession number PRJNA393501.

2.2.5 *De novo* genome assembly

FASTA and FASTQ files were extracted from base-called FAST5 files using *Poretools* (version 0.6.0) [239](Loman and Quinlan 2014). Raw nanopore reads were filtered for lambda

DNA by aligning to the *Enterobacteria phage* lambda reference genome (RefSeq assembly accession: GCF_000840245.1) using *Graphmap* [156](Sović et al.2016) with $-no-end2end$ parameter and retaining only unmappeds reads using *Samtools* [240](Li et al.2009). All reads obtained from the Delft and the Frankfurt CEN.PK113-7D stock cultures were assembled de novo using *Canu* (version 1.3) [151] with $-genomesize$ set to 12 Mbp. The assemblies were aligned using the *MUMmer* tool package: *Nucmer* with the $-maxmatch$ parameter and filtered for the best one-to-one alignment using *Delta-filter* [241]. The genome assemblies were visualised using *Mummerplot* [241] with the $-fat$ parameter. Gene annotations were performed using *MAKER2* annotation pipeline (version 2.31.9) using *SNAP* (version 2013-11-29) and *Augustus* (version 3.2.3) as *ab initio* gene predictors [242]. S288C EST and protein sequences were obtained from SGD (Saccharomyces Genome Database, `http://www.yeastgenome.org/`) and were aligned using *BLASTX* (BLAST version 2.2.28+) [243]. Translated protein sequence of the final gene model was aligned using BLASTP to S288C protein Swiss-Prot database. Custom made Perl scripts were used to map systematic names to the annotated gene names. Telomere repeat sequences (TEL07R of size 7306 bp and TEL07L of size 781 bp) from the manually curated and complete reference genome for *S. cerevisiae* S288C RefSeq assembly accession: GCA_000146045.2 obtained from SGD were aligned to the assembly as a proxy to assess completeness of each assembled chromosome. SGIDs for TEL07R and TEL07L are S000028960 and S000028887, respectively. The Tablet genome browser [244] was used to visualise nanopore reads aligned to the nanopore de novo assemblies. Short assembly errors in the Frankfurt assembly were corrected with *Nanopolish* (version 0.5.0) using default parameters [245] Two contigs, corresponding to chromosome XII, were manually scaffolded based on homology to S288C. To obtain the 2-μm native plasmid in CEN.PK113-7D, we aligned S288C's native plasmid to the 'unassembled' contigs file provided by *Canu* [151] and obtained the best aligned contig in terms of size and sequence similarity. Duplicated regions due to assembly difficulties in closing circular genomes were identified with *Nucmer* and manually corrected. *BWA-mem Li2010* was used to align Illumina reads to the scaffolded CEN.PK113-7D Frankfurt assembly using default parameters. *Pilon* [91] was then used to further correct assembly errors by aligning Illumina reads to the scaffolded Frankfurt assembly using correction of only SNPs and short indels ($-fix$ bases parameter) using only reads with a minimum mapping quality of 20 ($-minmq$ 20 parameter). Polishing with structural variant correction in addition to SNP and short indel correction was benchmarked, but not applied to the final assembly (see Additional File 1, Supporting Information in [207]).

2.2.6 Analysis of added information in the CEN.PK113-7D nanopore assembly

Gained and lost sequence information in the nanopore assembly of CEN.PK113-7D was determined by comparing it to the previous short-read assembly [141]. Contigs of at least 1 Kbp of short-read assembly were aligned to the nanopore CEN.PK113-7D Frankfurt assembly using the *MUMmer* tool package [241] using $-show-coords$ to extract alignment coordinates. For multimapped contigs, overlapping alignments of the same contig were collapsed and the largest alignment length as determined by *Nucmer* was used. Unaligned coordinates in the nanopore assembly were extracted and considered as added sequence. Added genes were retrieved by extracting the gene annotations in

these unaligned regions from the annotated nanopore genome; mitochondria and 2-μm plasmid genes were excluded for the lost sequence, unaligned sequences were obtained by aligning the contigs of the nanopore assembly to the short-read contigs of at least 1 kb using the same procedure as described above. Lost genes were retrieved by aligning the unaligned sequences to the short-read CEN.PK113-7D assembly with *BLASTN* (version 2.2.31 +) [243] and retrieving gene annotations. *BLASTN* was used to align DNA sequences of YHRCTy1-1, YDRCTy2-1, YILWTy3-1, YHLWTy4-1 and YCLWTy5-1 (obtained from the Saccharomyces Genome Database; SGIDs: S000007006, S000006862, S000007020, S000006991 and S000006831, respectively) as proxies for the location of two known groups of Ty-elements in *S. cerevisiae*, *Metaviridae* and *Pseudoviridae* [224], in the CEN.PK113-7D Frankfurt assembly. Non-redundant locations with at least a 2 Kbp alignment and an e-value of 0.0 as determined by *BLASTN* were then manually inspected.

2.2.7 Comparison of the CEN.PK113-7D assembly to the S288C genome

The nanopore assembly of CEN.PK113-7D and the reference genome of S288C (Accession number GCA_000146045.2) were annotated using the *MAKER2* pipeline described in the previous subsection 2.2.5. For each genome, a list of gene names per chromosome was constructed and compared strictly on their names to identify genes names absent in the corresponding chromosome in the other genome. The ORFs of genes identified as absent in either genome were aligned using *BLASTN* (version 2.2.31 +) to the total set of ORFs of the other genome and matches with an alignment length of half the query and with a sequence identity of at least 95% were listed. If one of the unique genes aligned to an ORF on the same chromosome, it was manually inspected to check if it was truly absent in the other genome. Merged ORFs and misannotations were not considered in further analysis. These alignments were also used to identify copies and homologues of the genes identified as truly absent in the other genome.

Gene ontology analysis was performed using the Gene Ontology term finder of SGD using the list of unique genes as the query set and all annotated genes as the background set of genes for each genome (Additional File 2A and 2C, Supporting Information). The ORFs of genes identified as present in S288C but absent in CEN.PK113-7D in previously made lists [141, 246] were obtained from SGD. The ORFs were aligned both ways to ORFs from SGD identified as unique to S288C in this study using BLASTN. Genes with alignments of at least half the query length and with a sequence identity of at least 95% were interpreted as confirmed by the other data set. In order to analyse the origin of genes identified as unique to S288C, these ORFs were aligned using BLASTN to 481 genome assemblies of various *S. cerevisiae* strains obtained from NCBI (Additional File 3, Supporting Information) and alignments of at least 50% of the query were considered. The top alignments were selected based on the highest sequence ID and only one alignment per strain was counted per gene.

2.2.8 Chromosome translocation analysis

Reads supporting the original and translocated genomic architectures of chromosomes III and VIII were identified via read alignment of raw nanopore reads. First, the translocation breakpoints coordinates were calculated based on whole-genome alignment of CEN.PK113-7D Delft assembly to S288C with *MUMmer*. A modified version of S288C was created con-

taining the normal architectures of all 16 chromosomes and the mitochondrial genome plus the translocated architecture of chromosomes III-VIII and VIII-III. The first nearest unique flanking genes at each breakpoint were determined using *BLASTN* (version 2.2.31 +) in reference to both S288C and the Delft CEN.PK113-7D nanopore assembly. Raw nanopore reads from CEN.PK113-7D Delft and Frankfurt were aligned to the modified version of S288C, and nanopore reads that spanned the translocation breakpoints as well as the unique flanking sequences were extracted. Supporting reads were validated by re-aligning them to the modified version of S288C using *BLASTN*.

2.3 Results

2.3.1 Sequencing on a single nanopore flow cell enables near-complete genome assembly

To obtain a complete chromosome level de novo assembly of *Saccharomyces cerevisiae* CENPK113-7D, we performed long-read sequencing on the ONT MinION platform. A fresh sample of CEN.PK113-7D was obtained from the original distributer Dr P. Kötter (further referred to as 'CEN.PK113-7D Frankfurt'), cultured in a single batch on YPD medium, and genomic DNA was extracted. CEN.PK113-7D Frankfurt was sequenced on a single R9 (FLO-MIN104) chemistry flow cell using the 2D ligation kit for the DNA libraries producing more than 49× coverage of the genome with an average read-length distribution of 10.0 Kbp (Fig. S1, Supporting Information) and an estimated error rate of 10% (Fig. S2, Supporting Information). We used *Canu* [151] to produce high-quality *de novo* assemblies using only nanopore data. Before correcting for misassemblies, the assembly contained a total of 21 contigs with an N50 of 756 Kbp (Table S1, Supporting Information). This represented a 19-fold reduction in the number of contigs and a 15-fold increase of the N50 in comparison to the short-read-only assembly of the first CEN.PK113-7D draft genome version [141] (Table 6.1).

Table 2.1: Comparison of 454/Illumina and nanopore de novo assemblies of CEN.PK113-7D. Summary of *de novo* assembly metrics of CEN.PK113-7D Delft and CEN.PK113-7D Frankfurt. For the short-read assembly, only contigs of at least 1 Kbp are shown [141]. The nanopore assembly of CEN.PK113-7D Delft is uncorrected for misassemblies while CEN.PK113-7D Frankfurt was corrected for misassemblies.

	Delft	Delft	Frankfurt
Data	Short-reads	Nanopore	Nanopore
Contigs (≥ 1 Kbp)	414	24	20
Largest contig	0.210 Mbp	1.08 Mbp	1.50 Mbp
Smallest Contig	0.001 Mbp	0.013 Mbp	0.085 Mbp
N50	0.048 Mbp	0.736 Mbp	0.912 Mbp
Total assembly size	11.4 Mbp	11.9 Mbp	12.1 Mbp

Most chromosomes of the nanopore *de novo* assembly are single contigs and are flanked by telomere repeats. Genome completeness was determined by alignment to the manually curated reference genome of the strain S288C RefSeq assembly accession: GCA_000146045.2 (see Table S2, Supporting Information in [207]). The two largest yeast chromosomes, IV and XII, were each split into two separate contigs, and two additional contigs (31 and

38 Kbp in length) corresponded to unplaced subtelomeric fragments. In particular, the assembly for chromosome XII was interrupted in the RDN1 locus—a repetitive region consisting of gene encoding ribosomal RNA estimated to be more than 1 Mbp long (Venema and Tollervey 1999). Since no reads were long enough to span this region, the contigs were joined with a gap.

Manual curation resolved chromosome III, chromosome IV and the mitochondrial genome. Chromosome IV was fragmented into two contigs at a locus of 11.5 Kbp containing two Ty-elements in S288C (coordinates 981 171-992 642). Interestingly, the end of the first contig and the start of the second contig had 8.8 Kbp of overlap (corresponding to the two Ty-elements) and one read spanned the repetitive Ty-elements and aligned to unique genes on the left and right flanks (EXG2 and DIN7, respectively). We therefore joined the contigs without missing sequence resulting in a complete assembly of chromosome IV. For chromosome III, the last ~27 Kbp contained multiple telomeric caps next to each other. The last ~10 Kbp had little to no coverage when re-aligning raw nanopore reads to the assembly (Fig. S3, Supporting Information). The coordinates for the first telomeric cap were identified, and the remaining sequence downstream was removed resulting in a final contig of size of 347 Kbp. The original contig corresponding to the mitochondrial genome had a size of 104 Kbp and contained a nearly identical ~20 Kbp overlap corresponding to start of the *S. cerevisiae* mitochondrial genome (i.e. origin of replication) (Fig. S4, Supporting Information). This is a common artefact as assembly algorithms generally have difficulties reconstructing and closing circular genomes [234, 247] The coordinates of the overlaps were determined with Nucmer [241] and manually joined resulting to a final size of 86 616 bp.

Overall, the final CEN.PK113-7D Frankfurt assembly contained 15 chromosome contigs, 1 chromosome scaffold, the complete mitochondrial contig, the complete 2-μm plasmid and two unplaced telomeric fragments, adding up to a total of 12.1 Mbp (Table 6.1 and Table S3, Supporting Information in [207]). Of the 16 chromosomes, 11 were assembled up until both telomeric caps, 4 were missing one of the telomere repeats and only chromosome X was missing both telomere repeats. Based on homology with S288C, the missing sequence was estimated not to exceed 12 kbp for each missing (sub)telomeric region. Furthermore, we found a total of 46 retrotransposons Ty-elements: 44 were from the *Pseudoviridae* group (30 Ty1, 12 Ty2, 1 Ty4 and 1 Ty5) and 2 from *Metaviridae* group (Ty3). The annotated nanopore assembly of CEN.PK113-7D Frankfurt is available at NCBI under the bioproject accession number PRJNA393501.

2.3.2 Comparison of the nanopore and short-read assemblies of CEN.PK113-7D

We compared the nanopore assembly of CEN.PK113-7D to a previously published version to quantify the improvements over the current state of the art [141]. Alignment of the contigs of the short-read assembly to the nanopore assembly revealed 770 Kbp of previously unassembled sequence, including the previously unassembled mitochondrial genome (Additional file 4A, Supporting Information in [207]). This gained sequence was relatively spread out over the genome (see Figure 2.1 A and B) and contained as much as 284 chromosomal gene annotations (Additional file 4B in [207]). Interestingly, 69 out of 284 genes had paralogues, corresponding to a fraction almost twice as high as the 13% found in the

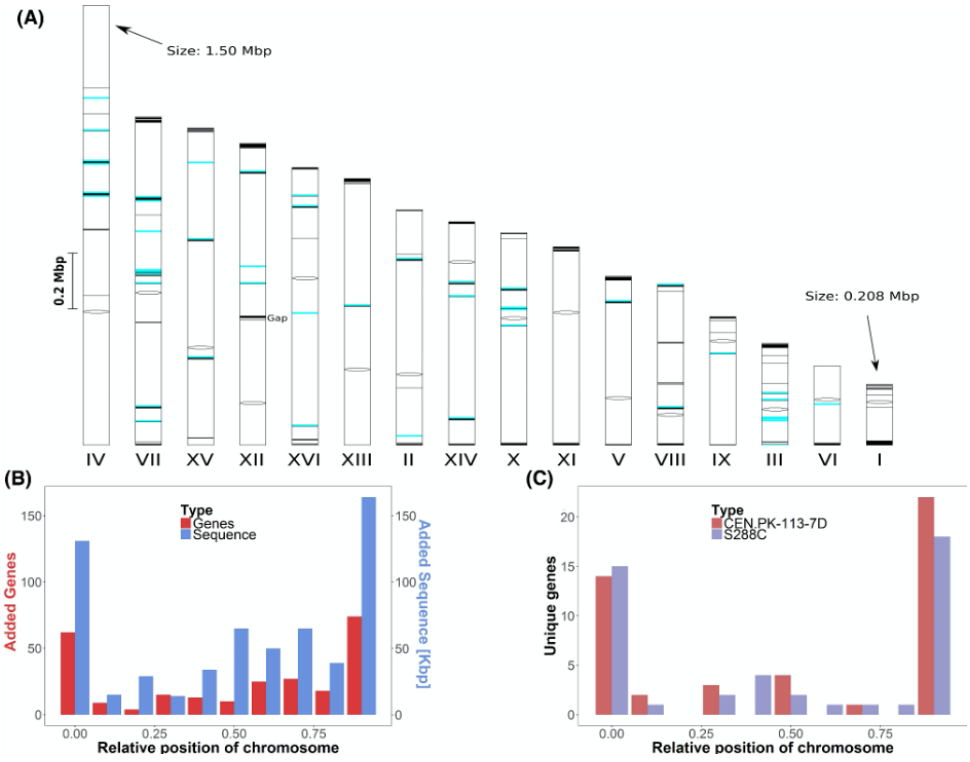


Figure 2.1: Overview of gained and lost sequence and genes in the CEN.PK113-7D Frankfurt nanopore assembly relative to the short-read CEN.PK113-7D assembly and to the genome of S288C. The two unplaced subtelomeric contigs and the mitochondrial DNA were not included in this figure. (A) Chromosomal location of sequence assembled in the nanopore assembly which was not assembled using short-read data. The 16 chromosome contigs of the nanopore assembly are shown. Chromosome XII has a gap at the RDN1 locus, a region estimated to contain more than 1 Mbp worth of repetitive sequence [247]. Centromeres are indicated by black ovals, gained sequence relative to the short-read assembly is indicated by black marks and 46 identified retrotransposon Ty-elements are indicated by blue marks. The size of all chromosomes and marks is proportional to their corresponding sequence size. In total, 611 Kbp of sequence was added within the chromosomal contigs. (B) Relative chromosome position of sequences and genes assembled on chromosome contigs of the nanopore assembly which were not assembled using short-read data. The positions of added sequence and genes were normalised to the total chromosome size. The number of genes (red) and the amount of sequence (cyan) over all chromosomes are shown per 10th of the relative chromosome size. (C) Relative chromosome position of gene presence differences between S288C and CEN.PK113-7D. The positions of the 45 genes identified as unique to CEN.PK113-7D and of the 44 genes identified as unique to S288C were normalised to the total chromosome size. The number of genes unique to CEN.PK113-7D (red) and to S288C (purple) are shown per 10th of the relative chromosome position.

whole genome of S288C [248]. Gene ontology analysis revealed an enrichment in the biological process of cell aggregation ($p = 9.30x10^{-4}$); in the molecular functions of mannose binding ($p = 3.90x10^{-4}$) and glucosidase activity ($p = 7.49x10^{-3}$); and in the cellular components of the cell wall ($p = 3.41x10^{-7}$) and the cell periphery component ($p = 5.81x10^{-5}$). Some newly assembled genes are involved in central carbon metabolism, such as PDC5. In addition, many of the added genes are known to be relevant in industrial applications including hexose transporters such as HXT genes and sugar polymer hydrolases such as IMA and MALx2 genes; several genes relevant for cellular metal homeostasis, such as CUP1-2 (linked to copper ion tolerance) and FIT1 (linked to iron ion retention); genes relevant for nitrogen metabolism in medium rich or poor in specific amino acids, including amino acid transporters such as VBA5, amino acid catabolism genes such as ASP3-4 and LEU2 and amino-acid limitation response genes such as many PAU genes; several FLO genes that are responsible for calcium-dependent flocculation; and various genes linked to different environmental stress responses, such as HSP genes increasing heat shock tolerance and RIM101 increasing tolerance to high pH.

To evaluate whether previously assembled sequences were missing in the nanopore assembly, we aligned the nanopore contigs to the short-read assembly [141]. Less than 6 Kbp of sequence of the short-read assembly was not present in the nanopore assembly, distributed over 13 contigs (Additional file 4C in [207]). Only two ORFs were missing: the genes BIO1 and BIO6 (Additional file 4D in [207]). Alignment of BIO1 and BIO6 sequences to the nanopore assembly showed that the right end of the chromosome I contig contains the first ~500 nt of BIO1. While BIO1 and BIO6 were present in the nanopore sequences, they are absent in the final assembly likely due to the lack of long-enough reads to resolve the repetitive nature of this subtelomeric region.

Overall, an additional 770 Kbp sequence containing 284 genes was gained, while 6 Kbp sequence containing two genes was not captured compared to the previous assembly. In addition, the reduction from over 700 to only 20 contigs clearly showed that the nanopore assembly is much less fragmented than the short-read assembly (Table 6.1).

2.3.3 Comparison of the nanopore assembly of CEN.PK113-7D to S288C

To identify unique and shared genes between CEN.PK113-7D and S288C, we compared annotations made using the same method for both genomes (Additional Files 2A and 2C in [207]). We identified a total of 45 genes unique to CEN.PK113-7D and 44 genes unique to S288C (Additional Files 2B and 2D [207]). Genes located in regions that had no assembled counterpart in the other genome were excluded: 20 for S288C and 27 for CEN.PK113-7D. Interestingly, the genes unique to either strain and genes present on different chromosomes were found mostly in the outer 10% of the chromosomes, indicating that the subtelomeric regions harbour most of the genetic differences between CEN.PK113-7D and S288C (Fig. 1C).

In order to validate the genes identified as unique to S288C, we compared them to genes identified as absent in CEN.PK113-7D in previous studies (Additional file 2D in [207], Table 2). A total of 25 genes of S288C were identified as absent in CEN.PK113-7D by array comparative genomic hybridisation analysis [246], and 21 genes were identified as absent in CEN.PK113-7D based on short-read WGS [141]. Of these genes, 19 and 10 respectively

were identified as genes in S288C by our annotation pipeline and could be compared to the genes we identified as unique to S288C. While 19 of these 29 genes were also absent in the nanopore assembly, the remaining 10 genes were fully assembled and annotated, indicating they were erroneously identified as missing (Table 2).

Table 2.2: Presence in the nanopore assembly of genes identified as absent in CEN.PK113-7D in previous research. For genes identified as absent in CEN.PK113-7D in two previous studies, the absence or presence in the nanopore assembly of CEN.PK113-7D is shown. A total of 25 genes were identified previously by aCGH [246] and 21 genes were identified by short-read genome assembly [141]. Genes that were not annotated by *MAKER2* in S288C could not be analysed. Genes with an alignment to genes identified as missing in the nanopore assembly of at least 50% of the query length and 95% sequence identity were confirmed as being absent, while those without such an alignment were identified as present. The presence of these genes was verified manually, which revealed the misanotation of YPL277C as YOR389W.

	Not analysed	**Absent in assembly**	**Present in assembly**
Daran-Lapujade *et al.*	*YAL064C-A, YAL066W, YAR047C, YHL046W-A, YIL058W, YOL013W-A*	*YAL065C, YAL067C, YBR093C, YCR018C, YCR105W, YCR106W, YDR038C, YDR039C, YHL047C, YHL048W, YNR070W, YNR071C* and *YNR074C*	*YAL069W, YDR036C, YDR037W, YJL165C, YNR004W,* and *YPL277C* (misannotated as *YOR389W*)
Nijkamp *et al.*	*Q0140, YDR543C, YDR544C, YDR545W, YIL046W-A, YLR154C-H, YLR156C-A, YLR157C-C, YLR159C-A, YOR029W* and *YOR082C*	*YBR093C, YCR040W, YCR041W, YDR038C, YDR039C* and *YDR040C*	*YDR036C, YHL008C, YHR056C* and *YLR055C*

In order to determine if the genes unique to S288C have homologues elsewhere in the genome of CEN.PK113-7D or if they are truly unique, we aligned the ORFs of the 44 genes identified as unique in S288C to the ORFs in the nanopore CEN.PK113-7D assembly. A total of 26 genes were completely absent in the CEN.PK113-7D assembly, while the remaining 18 genes aligned to between 1 and 20 ORFs each in the genome of CEN.PK113-7D with more than 95% sequence identity, indicating they may have close homologues or additional copies in S288C (Additional file 2D in [207]). Gene ontology analysis revealed no enrichment in biological process, molecular functions or cell components of the 26 genes without homologues in CEN.PK113-7D. Five genes without homologues were labelled as putative. However, there were many genes encoding proteins relevant for fitness under specific industrial conditions, such as PHO5 that is part of the response to phosphate scarcity, COS3 linked to salt tolerance, ADH7 linked to acetaldehyde tolerance, RDS1 linked to resistance to cycloheximide, PDR18 linked to ethanol tolerance and HXT17 that is involved in hexitol uptake (Additional file 2D in [207]). In addition, we confirmed

the complete absence of ENA2 and ENA5 in CEN.PK113-7D that are responsible for lithium sensitivity of CEN.PK113-7D [249].

In contrast, to determine if the genes unique to CEN.PK113-7D had homologues elsewhere in the genome of S288C or if they were truly unique, we aligned the ORFs of the 45 genes identified as unique in CEN.PK113-7D to the ORFs of S288C. A set of 16 genes were completely absent in S288C, while the remaining 29 aligned to between 1 and 16 ORFs each in the genome of S288C with more than 95% sequence (Additional File 2D) in [207]. Gene ontology analysis revealed no enrichment in biological processes, molecular functions or cell components of the 16 genes unique to CEN.PK113-7D without homologues. However, among the genes without homologues, a total of 13 were labelled as putative. The presence of an additional copy of IMA1, MAL31 and MAL32 on chromosome III was in line with the presence of the MAL2 locus that was absent in S288C. Interestingly, the sequence of MAL13, which belongs to this locus, was divergent enough from other MAL-gene activators not to be identified as homologue. Additionally, when performing the same analysis on the 27 genes on the two unplaced contigs of the CEN.PK113-7D assembly, 7 of them did not align to any gene of S288C with more than 95% sequence identity, indicating that these unplaced telomeric regions were highly unique to CEN.PK113-7D.

Since the genome of CEN.PK113-7D contained 45 ORFs that were absent in S288C, we investigated their origin by aligning them against all available *S. cerevisiae* nucleotide data at NCBI (Additional File 3 in [207]). For each ORF, we reported the strains to which they aligned with the highest sequence identity and the sequence identity relative to S288C in Additional File 2B in [207]. For most genes, several strains aligned equally well with the same sequence identity. For 13 ORFs, S288C is among the best matches, indicating these ORFs may come from duplications in the S288C genome. However, S288C was not among the best matches for 32 ORFs. In these, laboratory strain 'SK1' was among the best matches nine times, the west African wine isolate 'DBVPG6044' appeared eight times, laboratory strain 'W303' appeared seven times, the Belgian beer strain 'beer080' appeared three times and the Brazilian bioethanol strain 'bioethanol005' appeared three times. Interestingly, some grouped unique genes were most related to specific strains. For example, the unique genes identified on the left subtelomeric regions of chromosome XVI (YBL109W, YHR216W and YOR392) and of chromosome VIII (YJL225C and YOL161W) exhibited the highest similarity to DBVPG6044. Similarly, the right end of the subtelomeric region of chromosome III (YPL283W-A and YPR202) and of chromosome XI (YPL283W-A and YLR466W) were most closely related to W303.

Interestingly, the nanopore assembly revealed a duplication of LEU2, a gene involved in synthesis of leucine that can be used as an auxotrophy marker. In the complete reference genome of *S. cerevisiae* S288C, both LEU2 and NFS1 are unique, neighbouring genes located on chromosome III. However, gene annotations of the assemblies and raw nanopore reads supported additional copies of LEU2 and NFS1 in CEN.PK113-7D located on chromosome VII (Figure 2.2). The additional copy contained the complete LEU2 sequence but only ~0.5 kb of the 5′ end of NFS1. In CEN.PK113-7D and S288C, the LEU2 and NFS1 loci in chromosome III were located adjacent to Ty-elements. Two such Ty-elements were also found flanking the additional LEU2 and NFS1 loci in chromosome VII (Figure 2.2). It is likely that the duplication was the result of a translocation based on homology of the Ty-elements that resulted in local copy number increase during its strain development

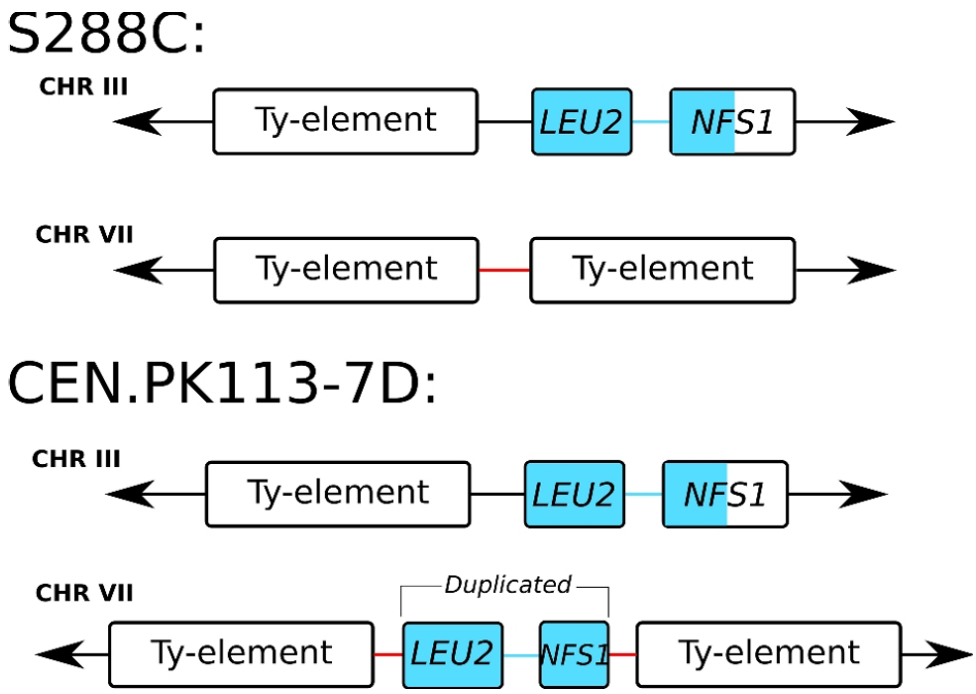


Figure 2.2: LEU2 and NFS1 duplication in chromosome VII of CEN.PK113–7D. The nanopore assembly contains a duplication of LEU2 and part of NFS1 in CEN.PK113–7D. In S288C, the two genes are located in chromosome III next to a Ty element. In CEN.PK113-7D, the two genes are present in chromosome III and in chromosome VII. The duplication appears to be mediated by Ty-elements. Note that the additional copy in chromosome VII is present in between two Ty-elements and contains only the first ~500 bp of NFS1. The duplication is supported by long-read data that span across the LEU2, NFS1, the two Ty-elements and the neighbouring flanking genes (not shown).

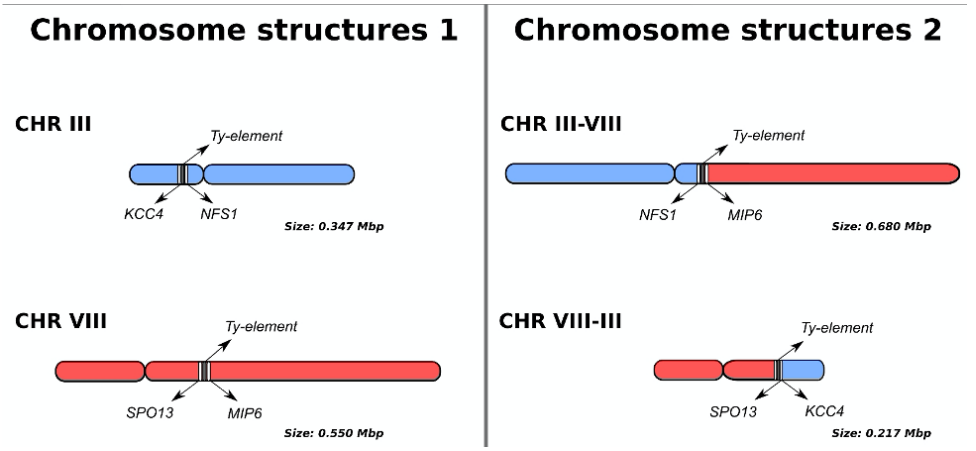


Figure 2.3: Overview of chromosome structure heterogeneity in CEN.PK113-7D Delft for CHRIII and CHRVIII that led to the misidentification of a fourth MAL locus in a previous short-read assembly study of the genome of CEN.PK113-7D. Nanopore reads support the presence of two chromosome architectures: the normal chromosomes III and VIII (left panel) and translocated chromosomes III-VIII and VIII-III (right panel). The translocation occurred in Ty-elements, large repetitive sequences known to mediate chromosomal translocations in Saccharomyces species [250]. Long reads are required to diagnose the chromosome architecture via sequencing: the repetitive region between KCC4 and NFS1 in chromosome III exceeds 15 Kbp, while the region between SPO13 and MIP6 in chromosome VIII is only 1.4 Kbp long. For the translocated architecture, the region from NFS1 to MIP6 in chromosome III-VIII exceeds 16 Kbp and the distance from SPO13 to KCC4 in chromosome VIII–III is nearly 10 Kbp.

program (Entian and Kötter 2007).

2.3.4 Long-read sequencing data reveals chromosome structure heterogeneity in CEN.PK113-7D Delft

CEN.PK113-7D has three confirmed MAL loci encoding genes for the uptake and hydrolysis of maltose: MAL1 on chromosome VIII, MAL2 on chromosome III and MAL3 on chromosome II (Additional file 2A in [207]). A fourth MAL locus was identified in previous research on chromosome XI based on contour-clamped homogeneous electric field electrophoresis (CHEF) and Southern blotting with a probe for MAL loci [141]. However, the nanopore assembly revealed no additional MAL locus despite the complete assembly of chromosome XI. The CEN.PK113-7D stock in which the fourth MAL locus was obtained from Dr P. Kötter in 2001 and stored at -80°since (further referred to as 'CEN.PK113-7D Delft'). In order to investigate the presence of the potential MAL locus, we sequenced CEN.PK113-7D Delft using nanopore MinION sequencing. Two R7.3 flow cells (FLO-MIN103) produced 55× coverage with an average read-length distribution of 8.5 Kbp and an R9 flow cell (FLO-MIN103) produced 47× coverage with an average read-length distribution of 3.2 Kbp (Fig. S1 in [207]). The error rate was estimated to be 13 (Fig. S4 [207]) after aligning the raw nanopore reads to the CEN.PK113-7D Frankfurt assembly. These reads were assembled into 24 contigs with an N50 of 736 Kbp (Table S1 in [207]).

Alignment of the assembly of CEN.PK113-7D Delft to the Frankfurt assembly showed evidence of a translocation between chromosomes III and VIII (Fig. S5, Supporting In-

formation [207]). The assembly thus suggested the presence of two new chromosomes: chromosomes III-VIII of size 680 Kbp and chromosome VIII-III of size 217 Kbp (Fig. 3). The translocation occurred between Ty-element YCLWTy2-1 on chromosome III and long-terminal repeats YHRCdelta5-7 on chromosome VIII. These repetitive regions were flanked by unique genes KCC4 and NFS1 on chromosome III and SPO13 and MIP6 on chromosome VIII (Fig. 3). Nanopore reads spanning the whole translocated or non-translocated sequence anchored in the unique genes flanking them were extracted for CEN.PK113-7D Delft and Frankfurt. A total of eight reads from CEN.PK113-7D Delft supported the translocated chromosome III-VIII architecture (largest read was 39 Kbp) and one 19 Kbp read supported the normal chromosome III architecture. For CEN.PK113-7D Frankfurt, we found only one read of size 23 Kbp that supported the normal chromosome III architecture but we found no reads that supported the translocated architectures. These data suggested that CEN.PK113-7D Delft was in fact a heterogeneous population containing cells with recombined chromosomes III and VIII and cells with original chromosomes III and VIII. As a result, in addition to the MAL2 locus on chromosome III, CEN.PK113-7D Delft harboured a MAL2 locus on recombined chromosome III-VIII. As the size of recombined chromosome III-VIII was close to chromosome XI, the MAL2 locus on chromosome III-VIII led to misidentification of a MAL4 locus on chromosome XI [141]. By repeating the CHEF gel and Southern blotting for MAL loci on several CEN.PK113-7D stocks, the MAL2 on the translocated chromosomes III-VIII was shown to be present only in CEN.PK113-7D Delft, demonstrating that there was indeed chromosome structure heterogeneity (Additional File 5 in [207]).

2.4 Discussion

In this study, we obtained a near-complete genome assembly of *Saccharomyces cerevisiae* strain CEN.PK113-7D using only a single R9 flow cell on ONT's MinION sequencing platform. Fifteen of the 16 chromosomes as well as the mitochondrial genome and the 2-μm plasmid were assembled in single, mostly telomere-to-telomere, contigs. This genome assembly is remarkably unfragmented, even when compared with other *S. cerevisiae* assemblies made with several nanopore technology flow cells, in which 18 to 105 chromosomal contigs were obtained [234, 237]. Despite the long-read lengths obtained by nanopore sequencing, the ribosomal DNA locus in chromosome XII could not be completely resolved. In practice, this would require reads exceeding 1 Mb in length, which current technology cannot yet deliver.

The obtained nanopore assembly is of vastly superior quality to the previous short-read-only assembly of CEN.PK113-7D that was fragmented into over 700 contigs [141]. In addition to the lesser fragmentation, the addition of 770 Kbp of previously unassembled sequence led to the identification and accurate placement of 284 additional ORFs spread out over the genome. These newly assembled genes showed overrepresentation for cell wall and cell periphery compartmentalisation and relate to functions such as sugar utilisation, amino acid uptake, metal ion metabolism, flocculation and tolerance to various stresses. While many of these genes were already present in the short-read assembly of CEN.PK113-7D, copy number was shown to be an important factor determining the adaptation of strains to specific growth conditions [233]. The added genes may therefore be very relevant for the specific physiology of CEN.PK113-7D under different industrial con-

ditions [233]. The ability of nanopore sequencing to distinguish genes with various similar copies is crucial in S. cerevisiae as homologues are frequent particularly in subtelomeric regions, and paralogues are widespread due to a whole genome duplication in its evolutionary history [248]. Besides the added sequence, 6 Kbp of sequence of the short-read assembly was not present in the nanopore assembly, mostly consisting of small unplaced contigs. Notably the absence of BIO1 and BIO6 in the assembly was unexpected, as it constituted a marked difference between CEN.PK113-7D and many other strains that enables biotin prototrophy [251]. Both genes were present in the nanopore reads, but were unassembled likely due to the lack of reads long enough to resolve this subtelomeric region (a fragment of BIO1 is located at the right end of chromosome I). Targeted long-read sequencing in known gaps of a draft assembly followed by manual curation could provide an interesting tool to obtain complete genome assemblies [252]. Alternatively, a more complete assembly could be obtained by maximising read length. The importance of read length is illustrated by the higher fragmentation of the CEN.PK113-7D Delft assembly compared to the Frankfurt one, which was based on reads with lower length distribution despite higher coverage and similar error rate (Table 6.1, Figs S1 and S5 in [207]). Read-length distribution in nanopore sequencing is highly influenced by the DNA extraction method and library preparation (Fig. S1 in [207]). The mitochondrial genome was completely assembled, which is not always possible with nanopore sequencing [234, 235, 237]. Even with identical DNA extraction and assembly methods, the mitochondrial genome cannot always be assembled, as illustrated by its absence in the assembly of CEN.PK113-7D Delft. Overall, the gained sequence in the nanopore assembly far outweighs the lost sequence relative to the previous assembly, and the reduction in number of contigs presents an important advantage.

The use of long-read sequencing enabled the discovery of a translocation between chromosomes III and VIII, which led to the misidentification of a fourth MAL locus on chromosome XI of CEN.PK113-7D [141]. Identification of this translocation required reads to span at least 12 Kbp due to the large repetitive elements surrounding the translocation breakpoints, explaining why it was previously undetected. While the translocation did not disrupt any coding sequence and is unlikely to cause phenotypical changes [253], there may be decreased spore viability upon mating with other CEN.PK strains. Our ability to detect structural heterogeneity within a culture shows that nanopore sequencing could also be valuable in detecting structural variation within a genome between different chromosome copies, which occurs frequently in aneuploid yeast genomes [209]. These results highlight the importance of minimal propagation of laboratory microorganisms to warrant genome stability and avoid heterogeneity that could at worst have an impact on phenotype and interpretation of experimental results.

The nanopore assembly of CEN.PK113-7D constitutes a vast improvement of its reference genome that should facilitate its use as a model organism. The elucidation of various homologue and paralogue genes is particularly relevant as CEN.PK113-7D is commonly used as a model for industrial *S. cerevisiae* applications for which gene copy number frequently plays an important role [209, 233]. Using the nanopore assembly as a reference for short-read sequencing of strains derived from CEN.PK113-7D will yield more complete and more accurate lists of SNPs and other mutations, facilitating the identification of causal mutations in laboratory evolution or mutagenesis experiments. Therefore, the

new assembly should accelerate elucidation of the genetic basis underlying the fitness of *S. cerevisiae* in various environmental conditions, as well as the discovery of new strain improvement strategies for industrial applications [254].

3

Alpaca: a kmer-based approach for investigating mosaic structures in microbial genomes

Microbial genomes are often mosaic: different regions can possess different evolutionary origins due to genetic recombination. Similarly, genome mosaicism can also arise from hybridization events from multiple species. The recent feasibility to assembling microbial genomes completely paired with existing sequencing datasets of large microbial populations enables researchers to investigate the potentially rich evolutionary history of a microbe at a much higher resolution. Here, I present Alpaca: a method that aims to investigate mosaicism using kmer similarity of large sequencing datasets. It does so by partitioning a given (complete) genome assembly into various sub-regions and comparing their similarity across a population of known genomes. The result is a high-resolution map of an entire genome and the most similar scoring clades across the given population.

3.1 Introduction

The ever-increasing availability of genomic data is enabling researchers to obtain novel insights about the genetic diversity and evolutionary history of various organisms [255, 256]. This is particularly important for microbes as they can originate from widely-diverse populations due lateral exchange of genetic information and hybridization of multiple genomes [137, 208, 209]. For example, horizontal gene transfer can lead to the introgression of novel sequences in nuclear chromosomes, aiding the adaptation for certain environments [255]. The resulting genome may thus be mosaic, meaning that different genomic segments may possess different evolutionary origins [257]. Genome hybridization—the joining of two or more genomes from different strains/species—can also lead to mosaic structures due to recombination of homologous segments followed by selection of those segments with fitness advantages [258]. As such, the genome of an individual microbe may be the result of a rich history of genetic adaptations after coming in contact with different populations

This chapter has been published in bioArxiv [213] and a component of BMC Genomics [212]

[255–258]. In other words, there can be multiple origins for the genetic content in a single genome.

We are currently at a crossroads of sequencing data-types. Long-read sequencing technologies (such as PacBio and Oxford Nanopore) are enabling researchers to *de novo* assemble complete individual microbial chromosomes, providing better insights in genome organization [146, 207, 259, 260]. Despite these advantages, the bulk of sequencing data is still stored as short-reads, requiring continue use them in population-wide studies [73, 261]. Our understanding of the mosaic structures in microbial population can thus benefit from a combination of both data-types, such as the structural basis provided by complete microbial genome assemblies from long-reads, along with the population diversity information in short-reads.

Here, I present *Alpaca*, a stand-alone method for investigating mosaic structures in high-quality microbial genome assemblies.

3.2 Method overview

The goal of *Alpaca* is to characterize mosaic structures in a (microbial) genome assembly. This is done by noting the change in similarity of local sequences throughout individual chromosomes relative to a given phylogenetic tree. Below we provide a detailed description of the foundations and implementation of *Alpaca*.

3.2.1 Alpaca foundations

A (microbial) genome with multiple evolutionary origins requires careful evaluation when compared to a reference population. Let R be the genome assembly of a microbial strain with only one chromosome. Given a set of evolutionarily similar genomes, $S = \{S_1, S_2, S_3, \cdots, S_n\}$, the evolutionary history of R can be inferred by constructing a phylogenetic tree and placing R in the context of all members in S. This requires a similarity calculation between any two genomes, traditionally done through multiple sequence alignments of core gene(s), whole-genome alignment, or whole-genome single-nucleotide polymorphisms [137, 208, 262–265]. However, if R formed through a recombination of two or more genomes in S, then the calculated similarities will fail to properly describe the multiple evolutionary origins of R since each calculation is either a global or local similarity.

The characterization of all local sequences in R could alternatively provide a better insight in its evolutionary history relative to the genomes in S. We can partition R as an ordered collection of subsequences, $R_p = \{r_1, r_2, r_3, \cdots, r_b\}$. Each partition, r_j, where $1 \leq j \leq b$, may be individual genes throughout R, where b is the total number of genes in the genome. Alternatively, R_p can represent non-overlapping sequences of length, l, such that $b = \lceil \frac{|R|}{l} \rceil$, and the concatenation of all subsequences in R_p would result in R. Calculating a local sequence similarity between each r_j and some genome in S, S_i, would thus describe changes in sequence similarity throughout R relative to S_i. More specifically, let S_{pi} be the partitioned version of S_i, such that $S_{pi} = \{s_1, s_2, s_3, \cdots, s_b\}$, and the sequences r_j and s_j are orthologous pairs (note that s_j can be an empty sequence if r_j is unique to R). The function $Similarity(R_p, S_{pi}, j)$ calculates the sequence similarity between r_j and s_j. Applying the similarity function through all b partitions would lead to a vector of similarity scores, which can be used to trace the change in sequence similarity throughout R relative to S_i.

The partitioning strategy described above can help pinpoint instances of mosaic structures in R. Let $MaxSim(r_j, S)$ be a function that returns a genome in S that yields the highest observed sequence similarity for a given r_j. If R has a linear evolutionary history (e.g. no recombination with other genomes), then the $MaxSim$ function will virtually always yield the same genome. In other words, the most similar genome throughout the subsequences in R will not change. More pragmatically, if a phylogenetic tree exists for all genomes in S, then $MaxSim$ will point to a specific evolutionary clade or lineage (e.g. the most similar clade or lineage does not change throughout the subsequences in R). However, if R does harbour mosaic structures, one would expect $MaxSim$ to yield different genomes from S, which would be particularly notable if those genomes consists of members from distant clades or lineages. A careful evaluation of the similarity vector can thus pinpoint neighbouring genes or subsequences that may derive from different evolutionary origins.

Although the partitioning strategy can theoretically help pinpoint instances of mosaic structures in R, it is not clear how to calculate $MaxSim(r_j, S)$, especially if S is a collection of short-read datasets. One approach is to perform a *de novo* genome assembly for each S_i, and partition it in such a way that a local sequence alignment can be perform to calculate the similarity against each r_j. However, this becomes problematic if S_i is a non-haploid genome and contains heterozygous alleles. As such, a *de novo* genome assembly will only result in a consensus representation without complete representation of all alleles present in S_i [137, 208, 212]. Consequently, the $Similarity$ function will yield erroneous sequence similarities of the orthologous pairs as it will fail to capture the heterozygous alleles. The same holds true if R is a non-haploid genome, as standard long-read assemblers generally yield consensus representations [212, 260]. A different approach is to identify variants by aligning reads from an S_i to each r_j. Although heterozygous single-nucleotide polymorphisms can be described, heterozygous structural variants are difficult to identify and represent, especially in short-read data [266]. Thus, the $Similarity$ function will still remain erroneous.

A calculation that considers heterozygous alleles between two orthologous pairs can provide a more accurate calculation for the $Similarity$ function. This can be done by considering all reads in both r_j and s_j, and not just their consensus representations. For example, if the original reads used to construct R were aligned back to the assembly, then all (partial) reads aligning to each r_j can be collected. From this set, all unique k-mers of length, k, can be identified. This procedure would thus construct a k-mer set containing all unique k-mers identified during the assembly of the sequence of r_j—implicitly capturing all alleles including potential heterozygous structural variants. Formally, let R_p now represent an ordered collection of sets where each r_j is a k-mer. Similarly re-defining S_{pi}, r_j and s_j are now the k-mer sets of the original orthologous sequences. The sequence similarity function, $Similarity(R_p, S_{pi}, j)$, can be re-defined through the Jaccard-Index, $Similarity(R_p, S_{pi}, j) = \frac{|r_j \cap s_j|}{|r_j \cup s_j|}$. Alternatively, a more biological metric closely following average nucleotide identity can be calculated using the MASH-distance, $Similarity(R_p, S_{pi}, j) = 1 - \frac{-1}{k} ln \frac{2j}{1+j}$, where j is the Jaccard-Index of r_j and s_j [177]. A vector of similarity scores describing the change in sequence similarity in R relative to some S_i that accounts for heterozygous alleles can thus be be obtained despite each S_i being a short-read dataset.

Although the genome partitioning produced described above assumes genomes with only a single chromosome, it is easier to adapt it for multi-chromosome organisms (e.g eukaryotic microbes). In this case, R and each S_i would consists of one or more chromosomal sequences. Their partitioned versions, R_p and S_{pi}, would be two-dimensional where each chromosome would contain an ordered collection of partitioned subsequences. The $MaxSim(r_j, S)$ would need be performed per chromosomal sequence as different chromosomes in the same nucleus can have different evolutionary histories.

In the next section, we describe how the partitioning strategy described above was implemented in a stand-alone method, *Alpaca*, that provides a custom visualization for the results.

3.2.2 Alpaca implementation

Alpaca will perform the partitioning procedure and similarity calculations described in the previous section through three major steps: (i) partitioning and storing a reference genome as a database, (ii) scoring the similarity of each sub-region for every sample in a given population, and (iii) a summary of all sub-regions with their top-scoring sample(s); ultimately providing insights about the potential mosaic structures in the given reference genome. The required inputs are a reference genome along with the original read-set used to construct it (representing to R in as described in the previous section), a collection of BAM-formatted files representing read-alignments to the provided reference from some collection of genomes (representing S), and a phylogenetic tree of the collection.

We provide a user-friendly, command-line implementation of these steps along with high-resolution and informative visualizations. Alpaca is written in the Scala programming language (https://www.scala-lang.org/) and packaged with a stand-alone binary distribution. Alpaca greatly benefits from parallelization (i.e. availability of multiple CPUs and multiple chromosome sequences) as it is implemented under a functional paradigm. We describe the major features of *Alpaca* below.

Alpaca database: genome partitioning of sub-regions

The first step is to create an Alpaca database for a given reference genome. The input is a FASTA-formatted assembly, a sorted BAM file of the native read alignments to the assembly (i.e. the same reads used to create the assembly). *Alpaca* will then iterate through each FASTA-entry and create non-overlapping sub-regions. Following the definitions in section 3.2.1, this corresponds to constructing the ordered collection of partitions, R_p, for each chromosome. By default, non-overlapping sub-regions of 2000 bp are created. Kmers (default size of 21) are then extracted for each sub-region in each chromosome, sampling from both the assembly and read-alignments independently using the *htsjdk* library. To minimize erroneous kmers, users can specify a minimum kmer count, or allow *Alpaca* to automatically detect the threshold based on alignment coverage. The output is a sub-directory describing all sub-regions and their corresponding kmer-sets.

Target comparison: computing sub-region similarity

The next step is to compute the similarity of all sub-regions in the reference genome against a target genome. The input is the path to the *Alpaca* database of the reference genome (see above), read alignments to the reference-genome's assembly as a sorted BAM

file, and the expected genome size of the target genome. Alpaca will then iterate through all sub-regions in the database and construct kmer-sets for the target genome. In other words, it will construct the ordered collection, S_{pi}, where each s_j is derived by sampling from the provided BAM file. The sequence similarity for each orthologous pair (corresponding to $Similarity(R_p, S_{pi}, j)$) is calculated via the Jaccard-Index as described in section section 3.2.1

The output is a tab-delimited file containing the coordinate of every sub-region in the database and the similarity score.

Population summary: ranking top-scoring samples

Comparing the sub-region similarities against a set of target genomes can provide insights for potential mosaic structures in the reference genome. The final step is thus to summarize every similarity score by ranking and retaining the top scoring target sample(s) for each sub-region. The input is the path to the *Alpaca* database of the subject genome and a tab-delimited file listing the path of the target comparison outputs from the previous step (see above). Alpaca will then iterate through each sub-region and retain the top-scoring targets. Note that there may be multiple top-scoring samples since different samples may possess the same similarity score. *Alpaca* will only retain the scores of sub-regions possessing a (configurable) number of top-scoring samples. The remaining samples have their similarity value set to zero, i.e. no similarity, to mitigate ambiguous sub-regions.The output is a tab-delimited file of every sub-region along with its top-scoring sample(s).

***Alpaca* layout: visualizing mosaic genomes**

If the used target population possesses labels (i.e. lineages, clades or species), then a high-resolution visualization for interpreting mosaic structures can be created (see Figure 3.1A). Using the population summary file, *Alpaca* will iterate through each chromosome and draw it as a sequence of rectangles representing individual sub-regions across the chromosome (see Figure 3.1A). The color of each sub-region represents the corresponding label of the top-scoring sample(s). Note that there can be multiple colors for each sub-region, whose proportions are based on the number of labels in the top-scoring samples. The corresponding proportion of a color for each sub-region is computed as the similarity score multiplied by the proportion of samples belonging to each label. To display unexplained similarity (i.e. sub-regions that are ambiguous or have low-scoring similarities), is drawn as white.

The resulting visualization (SVG-formatted) displays every sub-region and the corresponding similarity to assigned lineages or species, enabling users to identify instances of mosaic structures.

Tree-tracing: similarity across population-structures

The *Alpaca*-layout visualization is in context of lineages or species and therefore does not provide information about sub-region similarity to individual samples or sub-populations. If a (phylogenetic) tree in Newick-format (describing the population structure of different lineages or species) is provided, *Alpaca* will traverse the tree and hierarchically display the frequency of top-scoring samples and their corresponding (sub-) clades (see Figure 3.1B). Starting at the root-node, Alpaca will traverse through the tree and draw each branch with

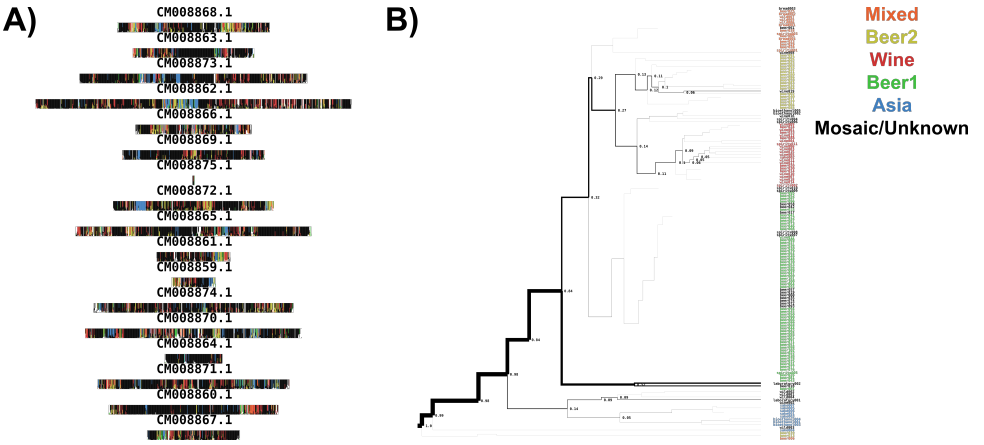


Figure 3.1: Summary visualizations of the mosaic profiles of the *S. cerevisiae* strain, CEN.PK113-7D, using 155 *S. cerevisiae* strains from *Gallone et al.* (A) Alpaca-layout figure: each rectangle is a chromosome from the assembly which is composed of a sequence of smaller rectangles (sub-regions) whose colors are based on the lineages (see colored legend in 1B) from the top-scoring samples within that sub-region. (B) Tree-tracing figure: the tree is a hierarchical clustering of the 155 strains and the width of the branches correspond to the overall frequency for which a given strain was a top-scoring sample for any sub-region. The color of the strain corresponds to the evolutionary lineage (Mixed, Beer2, Wine, Beer1, Asia, Mosaic/Unknown) as defined by *Gallone et al* [15].

a thickness corre-sponding to the proportion of a current node's children as top-scoring samples over the total sum.

3.3 Runtime and conclusion

We applied *Alpaca* to investigate potential mosaic profiles of the industrial *S. cerevisiae* strain, CEN.PK113-7D [207]. In general, the database creation took ~1.5 min with less than 1 gb of ram requiring ~57 mb of space with a single CPU using default parameters. We aligned short-read Illumina data from 155 *S. cerevisiae* strains from *Gallone et al.* [15] to the CEN.PK113-7D long-read assembly using *BWA-mem* [90] and computed genome similarities with *Alpaca* using default parameters. On average, target comparisons took ~2.5 hrs with less than 2 gb of ram on two CPUs. The results are summarized in Figure 3.1. To conclude, *Alpaca* can provide further insights of microbial assemblies by characterizing potential mosaic structures across sequencing datasets of microbial populations.

4

Chromosome level assembly and comparative genome analysis confirm lager-brewing yeasts originated from a single hybridization

The lager brewing yeast, S. pastorianus, is a hybrid between S. cerevisiae and S. eubayanus with extensive chromosome aneuploidy. S. pastorianus is subdivided into Group 1 and Group 2 strains, where Group 2 strains have higher copy number and a larger degree of heterozygosity for S. cerevisiae chromosomes. As a result, Group 2 strains were hypothesized to have emerged from a hybridization event distinct from Group 1 strains. Current genome assemblies of S. pastorianus strains are incomplete and highly fragmented, limiting our ability to investigate their evolutionary history.

To fill this gap, we generated a chromosome-level genome assembly of the S. pastorianus strain CBS 1483 from Oxford Nanopore MinION DNA sequencing data and analysed the newly assembled subtelomeric regions and chromosome heterozygosity. To analyse the evolutionary history of S. pastorianus strains, we developed Alpaca: a method to compute sequence similarity between genomes without assuming linear evolution. Alpaca revealed high similarities between the S. cerevisiae subgenomes of Group 1 and 2 strains, and marked differences from sequenced S. cerevisiae strains.

Our findings suggest that Group 1 and Group 2 strains originated from a single hybridization involving a heterozygous S. cerevisiae strain, followed by different evolutionary trajectories. The clear differences between both groups may originate from a population bottleneck caused by the isolation of the first pure cultures. Alpaca provides a computationally inexpensive

This chapter has been published in BMC Genomics [212]

method to analyse evolutionary relationships while considering non-linear evolution such as horizontal gene transfer and sexual reproduction, providing a complementary viewpoint beyond traditional phylogenetic approaches.

4.1 Introduction

The lager-brewing yeast *Saccharomyces pastorianus* is an interspecies hybrid between *S. cerevisiae* and *S. eubayanus*. Lager brewing emerged in the late middle ages and was carried out during winter months at temperatures between 8 and 15℃, followed by a prolonged maturation period referred to as lagering [2, 267]. While *S. cerevisiae* is a well-studied species frequently used in biotechnological processes [268], *S. eubayanus* was only discovered in 2011 and has thus far only been isolated from the wild [262]. Therefore, the ancestral *S. pastorianus* hybrid likely emerged from a spontaneous hybridization between an ale brewing *S. cerevisiae* yeast and a wild *S. eubayanus* contaminant, and took over lager brewing due to increased fitness under these conditions [262, 269, 270]. Indeed, laboratory-made *S. cerevisiae* x *S. eubayanus* hybrids demonstrated hybrid vigour by combining the fermentative capacity and sugar utilisation of *S. cerevisiae* and the ability to grow at lower temperatures of *S. eubayanus* [271, 272].

The genomes of *S. pastorianus* strains are highly aneuploid, containing 0 to 5 copies of each chromosome [137, 208–211, 269]. Between 45 and 79 individual chromosomes were found in individual *S. pastorianus* genomes, compared to a normal complement of 32 chromosomes in euploid *Saccharomyces* hybrids. The degree of aneuploidy of *S. pastorianus* is exceptional in the *Saccharomyces* genera, and likely evolved during its domestication in the brewing environment [209]. Nevertheless, two groups can be distinguished based on their genome organisation: Group 1 strains, which have approximately haploid *S. cerevisiae* and diploid *S. eubayanus* chromosome complements; and Group 2 strains, which have approximately diploid to tetraploid *S. cerevisiae* and diploid *S. eubayanus* chromosome complements [137, 208, 269, 273].

Group 1 and Group 2 strains in *S. pastorianus* were initially thought to have originated from two different hybridization events. Some lager-specific genes from Group 2 strains are absent in Group 1 strains, and the subtelomeric regions of Group 1 and Group 2 strains differ substantially [274, 275]. Based on these differences, Group 1 and Group 2 strains were hypothesized to have emerged from different independent hybridization events, involving a haploid S. cerevisiae for Group 1 strains and a higher ploidy *S. cerevisiae* strain for Group 2 strains [256, 269]. Indeed, crosses between *S. cerevisiae* and *S. eubayanus* strains with varying ploidies could be made in the laboratory, all of which performed well in the lager brewing process [276]. Comparative genome analysis between Group 1 and Group 2 strains revealed that there were more synonymous nucleotide differences in the *S. cerevisiae* subgenome than in the *S. eubayanus* subgenome [277]. As accumulation of synonymous mutations was presumed to equally affect both genomes, the authors hypothesized that Group 1 and 2 strains originated from two hybridizations, with a similar *S. eubayanus* parent and different *S. cerevisiae* parents.

More recent studies now support that Group 1 and Group 2 strains originated from the same hybridization event. Identical recombinations between the *S. cerevisiae* and *S. eubayanus* subgenomes were found at the ZUO1, MAT, HSP82 and XRN1/KEM1 loci in all analysed *S. pastorianus* strains [208, 210, 273], which did not emerge when such hy-

brids were evolved under laboratory conditions [278]. These conserved recombinations indicate that all *S. pastorianus* strains share a common *S. cerevisiae* x *S. eubayanus* hybrid ancestor, and that the differences between Group 1 and Group 2 strains emerged subsequently. Sequence analysis of ten *S. pastorianus* genomes revealed that the *S. cerevisiae* sub-genome in Group 1 strains is relatively homozygous, while Group 2 strains possess heterozygous sub-regions [208]. Moreover, heterozygous nucleotide stretches in Group 2 strains were composed of sequences highly similar to Group 1 genomes and of sequences from a different *S. cerevisiae* genome with a 0.5% lower sequence identity. As a result, the authors formulated two hypotheses to explain the emergence of Group 1 and Group 2 strains from a shared ancestral hybrid: (i) the ancestral hybrid had a heterozygous *S. cerevisiae* sub-genome, and Group 1 strains underwent a massive reduction of the *S. cerevisiae* genome content while Group 2 did not, or (ii) the ancestral hybrid had a homozygous Group 1-like genome and Group 2 strains were formed by a subsequent hybridization event of such a Group 1-like strain with another *S. cerevisiae* strain, resulting in a mixed *S. cerevisiae* genome content in Group 2 strains.

Since the exact *S. cerevisiae* and *S. eubayanus* ancestors of *S. pastorianus* are not available, the evolutionary history of *S. pastorianus* has so far been based on the sequence analysis using available *S. cerevisiae* and *S. eubayanus* reference genomes [208, 269]. However, these reference genomes are not necessarily representative of the original parental genomes of *S. pastorianus*. Although *S. pastorianus* genomes are available, they were sequenced with short-read sequencing technology [137, 208, 210, 211] preventing assembly of large repetitive stretches of several thousand base pairs, such as TY-elements or paralogous genes often found in *Saccharomyces* genomes [224]. The resulting *S. pastorianus* genomes assemblies are thus incomplete and fragmented into several hundred or thousand contigs [137, 208, 210, 211].

Single-molecule sequencing technologies can output reads of several thousand base pairs and span entire repetitive regions, enabling near complete chromosome-level genome assemblies of *Saccharomyces* yeasts [207, 234, 235, 237, 260, 279]. In addition to the lesser fragmentation, the assembly of regions containing repetitive sequences reveals large numbers of previously unassembled open reading frames, particularly in the sub-telomeric regions of chromosomes [207, 234, 279]. Sub-telomeric regions are relatively unstable [223], and therefore contain much of the genetic diversity between different strains [225, 233]. In *S. pastorianus*, notable differences were found between the sub-telomeric regions of Group 1 and Group 2 strains [274, 275], which could be used to understand their origin. Moreover, repetitive regions are enriched for genes with functions determining the cell's interaction with its environment, such as nutrient uptake, sugar utilization, inhibitor tolerance and flocculation [228–231]. As a result, the completeness of sub-telomeric regions is critical for understanding genetic variation and evolutionary relationships between strains, as well as for understanding their performance in industrial applications [207, 225, 233].

Here, we used Oxford Nanopore MinION sequencing to obtain a chromosome-level assembly of the Group 2 *S. pastorianus* strain CBS 1482 and analysed the importance of new-found sequences relative to previous genome assemblies, with particular focus on industrially-relevant subtelomeric gene families. As the CBS 1483 genome contains multiple non-identical copies for many chromosomes, we analyse structural and sequence-level heterozygosity using short- and long-read data. Moreover, we developed a method to in-

vestigate the evolutionary origin of *S. pastorianus* strains relative to a large dataset of *S. cerevisiae* and *S. eubayanus* genomes, including an isolate of the Heineken A-yeast® lineage which was isolated by dr. Elion in 1886 and is still used in beer production today.

4.2 Methods

4.2.1 Yeast strains, cultivation techniques and genomic DNA extraction

Saccharomyces strains used in this study are indicated in Table 3. *S. pastorianus* strain CBS 1483, *S. cerevisiae* strain S288C and *S. eubayanus* strain CBS 12357 were obtained from the Westerdijk Fungal Biodiversity Institute (`http://www.westerdijkinstitute.nl/`). *S. eubayanus* strain CDFM21L.1 was provided by Prof. Feng-Yan Bai. An isolate from the *S. pastorianus* Heineken A-yeast® lineage (Hei-A) was obtained from HEINEKEN Supply Chain B.V., Zoeterwoude, The Netherlands. All strains were stored at -80℃ in 30% glycerol (vol/vol). Yeast cultures were inoculated from frozen stocks into 500-mL shake flasks containing 100 mL liquid YPD medium (containing 10g L-1 yeast extract, 20g L-1 peptone and 20g L-1 glucose) and incubated at 12℃ on an orbital shaker set at 200 rpm until the strains reached stationary phase with an OD660 between 12 and 20. Genomic DNA was isolated using the Qiagen 100/G kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions and quantified using a Qubit® Fluorometer 2.0 (ThermoFisher Scientific, Waltham, MA).

Table 4.1: *Saccharomyces* strains used in this study. For strains of the reference dataset, please refer to their original publication [15, 265].

Name	Species	Description	Reference
CBS 1483	*S. pastorianus*	Group 2	[137]
CBS 2156	*S. pastorianus*	Group 2	[208]
WS 34/70	*S. pastorianus*	Group 2	[211]
Heineken A-yeast®	*S. pastorianus*	Group 2	This study
CBS 1503	*S. pastorianus*	Group 1	[208]
CBS 1513	*S. pastorianus*	Group 1	[210]
CBS 1538	*S. pastorianus*	Group 1	[208]
S288C	*S. cerevisiae*	Laboratory strain	[70]
CEN.PK113-7D	*S. cerevisiae*	Laboratory strain	[207]
CBS 7539	*S. cerevisiae*	Ale brewing strain	[73]
CBS 1463	*S. cerevisiae*	Ale brewing strain	[73]
A81062	*S. cerevisiae*	Ale brewing strain	[276]
CBS 1171	*S. cerevisiae*	Ale brewing strain	[73]
CBS 6308	*S. cerevisiae*	Ale brewing strain	[73]
CBS 1487	*S. cerevisiae*	Ale brewing strain	[73]
CBS 12357	*S. eubayanus*	Patagonian Isolate	[262]
CDFM21L.1	*S. eubayanus*	Himalayan isolate	[280]

4.2.2 Short-read Illumina sequencing

Genomic DNA of CBS 1483 and CDFM21L.1 was sequenced on a HiSeq2500 sequencer (Illumina, San Diego,CA) with 125 bp paired-end reads with an insert size of 550 bp using PCR-free library preparation by Keygene(Wageningen, The Netherlands). Genomic DNA of theHeineken A-yeast® isolate Hei-A was sequenced in house on a MiSeq sequencer (Illumina) with 300 bp paired-end reads using PCR-free library preparation. All Illumina sequencing (see Additional file 9: Table S1 in [212]) data are available at NCBI (`https://www.ncbi.nlm.nih.gov/`) under the bioproject accession number PRJNA522669.

4.2.3 Oxford nanopore minION sequencing and basecalling

A total of four long-read genomic libraries of CBS 1483 were created using different chemistries and flow cells: one library using 2D-ligation (Sequencing Kit SQK-MAP006)with a R7.3 chemistry flow cell (FLO-MIN103); two libraries using 2D-ligation (Sequencing Kit SQK-NSK007) witht two R9 chemistry flow cells (FLO-MIN105); and one library using 1D-ligation (Sequencing Kit SQK-LASK108) with a R9 chemistry flow cell (FLO-MIN106). All libraries were constructed using the same settings as previously described [207] and reads were uploaded and basecalled using the Metrichor desktop agent (`https://metrichor.com/s/`). All sequencing data (see Additional file 9: Table S1 in [212]) are avail-able at NCBI (`https://www.ncbi.nlm.nih.gov/`) under the BioProject accession number PRJNA522669.

4.2.4 *De novo* genome assembly

The genome of CBS 1483 was assembled de novo using only the long-read sequencing data generated in thisstudy. The assembly was generated using *Canu* [151], polished using Pilon [91] and annotated using MAKER2[92], as previously described [24] with some modifications: Pilon (version 1.22) was only used to polish sequencing errors in the long-read-only de novo assembly, and Minimap2 [93] (version 2.7) was used as the long-read aligner to identify potential mis-assemblies and heterozygous structural variants, which were visualized using Ribbon [94]. The resulting assembly was manually curated: (i) a contig of 24 Kbp comprised entirely of "TATATA" sequence was discarded; (ii) three contigs of592, 465, and 95 Kbp (corresponding to the rDNA locus of the *S. cerevisiae* subgenome) and complete sequence up and downstream of this locus were joined with a gap;(iii) four contigs corresponding to *S. cerevisiae* chromosome I (referred to asScI) were joined without a gap intoa complete 208 Kbp chromosome assembly (Fig.2a); (iv) two contigs corresponding toScXIV were joined with a gap (Fig.2d); and (v) 23 Kbp of overlapping sequence from the mitochondrial contig corresponding to the origin of replication was identified with *Nucmer* [95] and manually removed when circularizing the contig, leading to the complete a final size of 69 Kbp. The assembled genomes are available at NCBI (`https://www.ncbi.nlm.nih.gov/`) under the bioproject accession number PRJNA522669. Gene annotations are available in Additional file1 A.

4.2.5 Comparison between ONT-only and Illumina-only genome assembly

Gained and lost sequence information in the long-read assembly of CBS 1483 was determined by comparing it to the previous short-read assembly [137], as previously described [207] with the addition of using minimum added sequence length of 25 nt.

4.2.6 FLO gene analysis

We used *Tandem Repeat Finder* (version 4.09) [281] with recommended parameters to identify tandem repeat sequences in FLO1 (SGDID:S000000084), FLO5 (SGDID:S000001254), FLO8 (SGDID:S000000911), FLO9 (SGDID:S000000059), FLO10 (SGDID:S000001810), and FLO11 (SGDID:S000001458) of *S. cerevisiae* strain S288C [219] as well as in FLO1, FLO5, FLO8, FLO9, FLO10 and FLO11 of *S. eubayanus* strain CBS 12357 [271]. The resulting tandem repeat sequences were then used as proxies to characterize FLO genes in our assembly of CBS 1483, in a previously generated assembly of *S. cerevisiae* strain CEN.PK113-7D [207] and the Lg-FLO1 genes previously described in *S. cerevisiae* strain CMBSVM11 (GenBank HM358276) and *S. pastorianus* strain KBY001 (GenBank D89860.1) [282, 283]. BLASTN (version 2.2.31+) [243] was then used to align the tandem sequences to each FLO gene. The alignments were further processed via an in-house script in the Scala programming language to identify repeat clusters by requiring a minimum alignment coverage of 0.5 and a maximum gap between two repeats of 3x times the repeat sequence length. The total number of copies was estimated by dividing the total size of the cluster by the repeat sequence length.

4.2.7 Intra-chromosomal heterozygosity

Sequence variation was identified by aligning the short-read Illumina reads generated in this study to the long-read-only assembly with *BWA-mem* [90] (version 0.7.12) and calling variants with *Pilon* [91] using the $--fix$"$bases$", "$local$" and $--diploid$ parameters. To restrict false positive calls, SNPs were disregarded within 10 Kbp of the ends of the chromosomes, if minor alleles had a frequency below 15% allele frequency, and if the coverage was below 3 reads.

Copy number variation for all chromosomes were estimated by aligning all short-reads to the ONT-only assembly. Reads were trimmed of adapter sequences and low-quality bases with *Trimmomatic* [284] (version 0.36) and aligned with *BWA-mem*. The median coverage was computed using a non-overlapping window of 100 nt, copy number was determined by comparing the coverage to that of the chromosome with the smallest median coverage. Additionally, copy number variation at the gene-level was also investigated based on whether the coverage of an individual gene significantly deviated from the coverage of the surrounding region. First, we defined contiguous chromosomal sub-regions with fixed copy number (Table S2 in [212]). The mean and standard deviation of coverages of these sub-regions were then computed using long-read-only alignments. Mean coverages of every gene was then computed and an uncorrected Z-test was performed by comparing a gene's mean coverage and the corresponding mean and standard deviation of the pre-defined sub-region that the gene overlapped with.

4.2.8 Similarity analysis and lineage tracing of S. pastorianus sub-genomes using *Alpaca*

We developed *Alpaca* [213] to investigate non-linear ancestry of a reference genome based on large sequencing datasets. Briefly, Alpaca partitions a reference genome into multiple sub-regions, each reduced to a kmer set representation. Sequence similarities of the sub-regions are then independently computed against the corresponding sub-regions in a collection of target genomes. Non-linear ancestry can therefore be inferred by tracing the

population origin of the most similar genome(s) in each sub-region. Detailed explanation *Alpaca* can be found in our method description [213].

Alpaca (version 1.0) was applied to the long-read CBS 1483 genome assembly to investigate the similarity of sub-regions from both sub-genomes to previously defined population lineages. For partitioning the CBS 1483 genome into sub-regions, we used a kmer size of 21 and a sub-region size of 2 Kbp and used the short-read Illumina data of CBS 1483 produced in this study to assure accurate kmer set construction. For investigating mosaic structures in the *S. cerevisiae* subgenome, we used 157 brewing-related *S. cerevisiae* genomes (project accession number PRJNA323691) which were subdivided in six major lineages: Asia, Beer1, Beer2, Mixed, West-Africa, Wine and Mosaic *Gallone2016*. For the *S. eubayanus* subgenome, we used 29 available genomes (project accession number PRJNA290017) which were subdivided in three major lineages: Admixed, Patagonia-A, and Patagonia-B [265]. Raw-reads of all samples were trimmed Trimmomatic and filtered reads were aligned to CBS 1483 genome using *BWA-mem*. *Alpaca* was also applied to several *Saccharomyces* genomes to investigate evolutionary similarities and differences between Group 1 and Group 2 *S. pastorianus* genomes. We used Group 1 strains CBS 1503, CBS 1513, and CBS 1538, and Group 2 strains CBS 2156 and WS34/70 (project accession number PRJDB4073) [208]. As a control, eight *S. cerevisiae* genomes were analysed: ale strains CBS 7539, CBS 1463, CBS 1171, CBS 6308, and CBS 1487 (project accession number PRJEB13017) [73] and A81062 (project accession number PRJNA408119) [276], and laboratory strains CEN.PK113-7D (project accession number PRJNA393501) [207] and S288C (project accession number PRJEB14774) [237]. Similarly, raw-reads for all strains were trimmed with *Trimmomatic* and aligned to the long-read CBS 1483 genome assembly using *BWA-mem*. Partitioning of the additional *S. pastorianus* and *S. cerevisiae* genomes with *Alpaca* was performed by deriving kmer sets from read-alignments only, assuring direct one-to-one comparison of all sub-regions across all genomes. Kmer size of 21 and sub-region size of 2 Kbp were used. The *S. cerevisiae* and *S. eubayanus* sequencing data were used to identify potential mosaic structures in these genomes. Lastly, *S. cerevisiae* and *S. eubayanus* strains were subdivided into subpopulations according to previously defined lineages [15, 265]. *MASH* (version 2.1) [177] was then used to hierarchically cluster each genome based on their *MASH* distance using kmer size of 21, sketch size of 1,000,000, and minimum kmer frequency of 2. The resulting trees were used as population reference trees for *Alpaca* [213].

4.3 Results

4.3.1 Near-complete haploid assembly of CBS 1483

We obtained 3.3 Gbp of whole genome sequencing data of the *Saccharomyces pastorianus* strain CBS 1483 using 4 flow cells on Oxford Nanopore Technology's MinION platform. Based on a genome size of 46 Mbp accounting for all chromosome copy numbers, the combined coverage was 72x with an average read length of 7 Kbp (Additional file 2: Figure S1 in [212]). We assembled the reads using *Canu* [151] and performed manual curation involving circularization of the mitochondrial DNA, scaffolding of ScXII (chromosome XII of the *S. cerevisiae* sub-genome) and resolution of assembly problems due to inter- and intra-chromosomal structural heterozygosity in ScI and ScXIV (Figure 4.1). Assembly er-

Table 4.2: Length and gaps of each assembled chromosome of the *S. cerevisiae* and *S. eubayanus* subgenome in the *de novo* assembly of Group 2 *S. pastorianus* strain CBS 1483. The mitochondrial DNA is not shown.

***S. cerevisiae* sub-genome**			***S. eubayanus* sub-genome**		
Contigs/Scaffold	**Size**	**Gaps**	**Contigs/Scaffold**	**Size**	**Gaps**
ScI	208,794	0	SeI	183,365	0
ScII	812,290	0	SeII	1,284,912	0
ScIII	0	0	SeIII	311,639	0
ScIV	1,480,484	0	SeIV	995,872	0
ScV	590,259	0	SeV	580,717	0
ScVI	263,951	0	SeVI	268,897	0
ScVII	862,436	0	SeVII	1,048,199	0
ScVIII	547,874	0	SeVIII	813,607	0
ScIX	426,203	0	SeIX	413,986	0
ScX	772,632	0	SeX	698,708	0
ScXI	662,864	0	SeXI	658,371	0
ScXII	1,128,411	2	SeXII	1,043,408	0
ScXIII	872,991	0	SeXIII	966,749	0
ScXIV	783,474	0	SeXIV	765,784	1
ScXV	1,060,500	0	SeXV	754,183	0
ScXVI	926,828	0	SeXVI	788,293	0
Unplaced	36,198	0	Mitochondria	68,765	0

rors were corrected with Pilon [91]using paired-end Illumina reads with 159x coverage. We obtained a final assembly of 29 chromosome contigs, 2 chromosome scaffolds, and the complete mitochondrial contig leading to a total size of 23.0 Mbp (Figure 4.2 and Table 5.2). The assembly was remarkably complete: of the 31 chromosomes (in CBS 1483 ScIII and SeIII recombined into a chimeric SeIII-ScIII chromosome [137], 29 were in single contigs; 21 of the chromosomes contained both telomere caps; 8 contained one of the caps; and 2 were missing both caps. Some chromosomes contain sequence from both parental subgenomes due to recombinations; those chromosomes were named SeIII-ScIII, SeVII-ScVII, ScX-SeX, SeX-ScX and SeXIII-ScXIII, in accordance with previous nomenclature [137]. Annotation of the assembly resulted in the identification of 10,632 genes (Additional file 1a). We determined chromosome copy number based on coverage analysis of short-read alignments to the genome assembly of CBS 1483 (Figure 4.2 and Additional file 3: Figure S2 in [212]).

4.3.2 Comparison between Oxford nanopore minION and Illumina assemblies

In order to compare our novel long-read assembly of CBS 1483 to the previous assembly generated using short-read data, we aligned contigs of CBS 1483 from van den Broek *et al.* [137] to our current long-read assembly, revealing a total 1.06 Mbp of added sequence. The added sequence overlapped with 323 ORFs (Additional file 1b in [212]). Conversely, align-

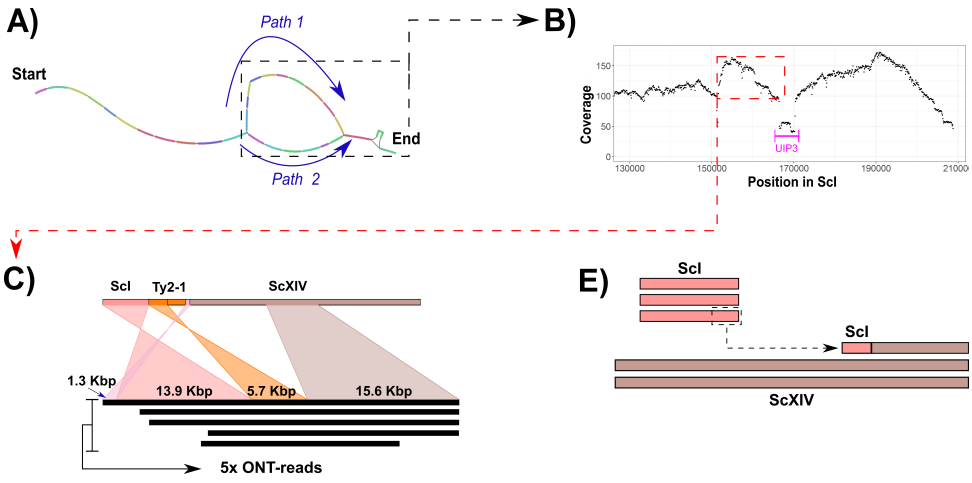


Figure 4.1: Structural heterozygosity within multiple copies of the *S. cerevisiae* chromosome I of CBS 1483. (A) Layout of *S. cerevisiae* chromosome I in the assembly graph. Paths 1 and 2 (blue text and arrows) represent alternative contigs in the right-end of the chromosome—the gene UIP3 is deleted in path 2. (B) Sequencing coverage of long-read alignments of CBS 1483 in the right-end of chromosome I after joining path 1 and discarding path 2. The location of the UIP3 gene is indicated. c Alignment overview of five raw long-reads supporting the introgression of a ~14 Kbp in chromosome I (pink colour) to a region at the right-end of chromosome XIV (brown colour) in the *S. cerevisiae* sub-genome. The additional alignments (pink and orange) are alignments to computationally confirmed Ty-2 repetitive elements. d Schematic representation of the two chromosome architectures of *S. cerevisiae* chromosome XIV (brown colour) due to translocation of an additional copy of the right arm of chromosome I (salmon colour) to the left arm of chromosome XIV.

ing the long-read assembly to the van den Broek *et al.* assembly revealed that only 14.9 Kbp of sequence were lost, affecting 15 ORFs (Additional file 1c in [212]). Gene ontology analysis of the added genes showed enrichment of several biological processes, functions, and components such as flocculation ($p = 7.44x10^{-3}$ as well as transporter activity for several sugars including mannose, fructose and glucose ($p \leq 1.5x10^{-5}$) (Additional file in 1d in [212]). Among the added genes were various members of subtelomeric gene families such as the FLO, SUC, MAL, HXT and IMA genes (Additional file 1e in [212]). Due to their role in the brewing-relevant traits such as carbohydrate utilization and flocculation, the complete assembly of subtelomeric gene families is crucial to capture different gene versions and copy number effects.

The assembly of CBS 1483 contained 9 MAL transporters, which encode for the ability to import maltose and maltotriose [285–287], constituting 85% of fermentable sugar in brewer's wort [288]. The *S. cerevisiae* subgenome harboured ScMAL31 on ScII, ScMAL11 on ScVII and on SeVII-ScVII, and ScMAL41 on ScXI (Additional File 1B and 1E in [212]). However, the ScMAL11 gene, also referred to as AGT1, was truncated, and there was no ScMAL21 gene due to the complete absence of ScIII, as reported previously [137, 274]. In the *S. eubayanus* subgenome, MAL31-type transporter genes were found in SeII, SeV, and SeXIII-ScXIII, corresponding to the location of the *S. eubayanus* transporter genes SeMALT1, SeMALT2 and SeMALT3, respectively [271]. In addition, a MAL11-like transporter was found on SeXV. Consistently with previous reports, no MTY1-like maltotriose transporter was found in CBS 1483 [137]. Due to the absence of MTY1 and the truncation of ScMAL11, maltotriose utilisation is likely to rely on the SeMAL11 transporter in CBS 1483. Indeed, a MAL11-like transporter was recently shown to confer maltotriose utilisation in an *S. eubayanus* isolate from North Carolina [289].

The assembly also contained 14 FLO genes encoding flocculins which cause cell mass sedimentation upon completion of sugar consumption [231, 290, 291]. The heavy flocculation of *S. pastorianus* cells simplifies biomass separation at the end of the brewing process, and resulted in their designation as bottom-fermenting yeast [292]. Flocculation is mediated by flocculins: lectin-like cell wall proteins which effect cell-to-cell adhesion. In CBS 1483, we identified 12 flocculin genes, in addition to two FLO8 transcriptional activators of flocculins (Additional File 1E). Flocculation intensity has been correlated to the length of flocculin genes [293–295]. Specifically, increased length and number of tandem repeats within the FLO genes caused increased flocculation [295, 296]. We therefore analysed tandem repeats in *S. cerevisiae*, *S. eubayanus* and *S. pastorianus* genomes and found that most FLO genes contain a distinct repeat pattern: two distinct, adjacent sequences each with variable copy number (Table 4.3). The repeats in FLO1, FLO5, and FLO9 of the *S. cerevisiae* strain S288C have the same repeats of 135 bp and 15 bp; while repeats are of 189 bp and 15 bp for FLO10 and of 132 bp and 45 bp for FLO11. The same repeat structures can be found in the *S. eubayanus* strain CBS 12357 as FLO1, FLO5, and FLO9 contain repeats of 156 and 30 bp; although we were unable to find clear repeat patterns for FLO10 and FLO11 in this genome. In S. pastorianus CBS 1483, the repeat lengths of FLO genes corresponded to the subgenome they were localized in (Table 4.3). Compared to the non-flocculent S288C and CBS 12357 strains, FLO genes were systematically shorter in CBS 1483, contrasting with available theory [290–295, 295–298]. The intense flocculation phenotype of S. pastorianus was previously attributed to a gene referred to as LgFLO1 [282, 297, 298]. However, align-

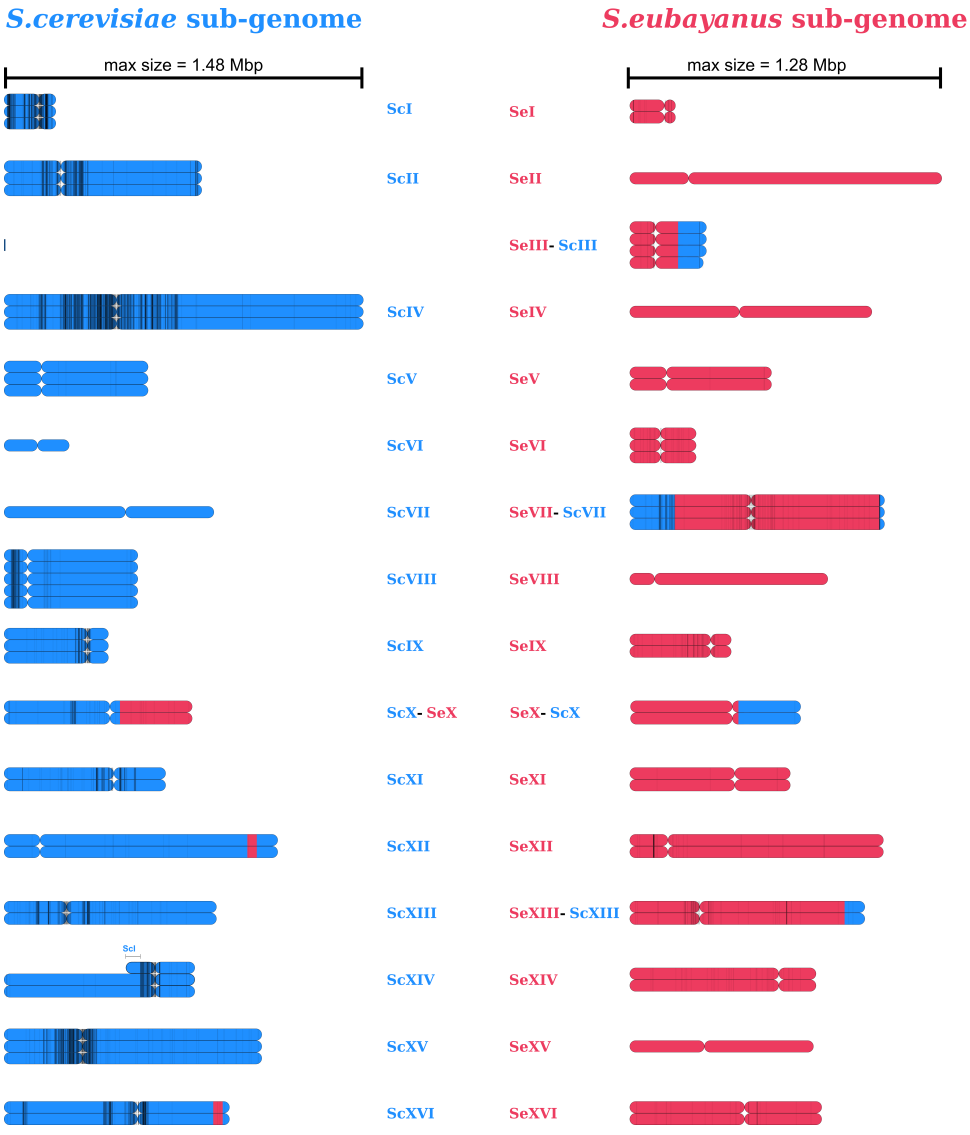


Figure 4.2: Overview of the long-read-only de novo genome assembly of the *S. pastorianus* strain, CBS 1483. For each chromosome, all copies are represented as coloured rectangles. Genomic material originating from *S. cerevisiae* (blue) and from *S. eubayanus* (red) are shown, and the position of the centromere is indicated by the constricted position within each rectangle. Heterozygous SNP calls are represented as vertical, black lines and are drawn with transparency to depict the density of SNP calls in a given region. Underlying chromosome copy number data and the list of heterozygous SNPs is available in Additional file 3: Figure S2 and Additional file 1 F in [212].

Table 4.3: Tandem repeat analysis in FLO genes. We found seven repeat sequences when analysing flocculation genes FLO1, FLO5, FLO9, FLO10, and FLO11 in *S. cerevisiae* (S288C) and *S. eubayanus* (CBS 12357) genomes. These sequences are referred to as sequence A (135 nt), B (15 nt), C (189 nt), D (45 nt), E (132 nt), F (156 nt), and G (30 nt). We used these sequences to analyse the copy numbers of each repeat within all FLO genes in our long-read-only assembly of CBS 1483 using the long-read-only S288C assembly as a control. Their respective copy numbers are shown below.

Species	(Sub)genome	Gene	Gene size (nt)	A	B	C	D	E	F	G
S. cerevisiae										
	S288C									
		FLO1	4614	18.0	9.4	-	-	-	-	-
		FLO5	3228	8.0	9.6	-	-	-	-	-
		FLO9	3969	13.0	8.3	-	-	-	-	-
		FLO10	3510	-	3.8	4.4	-	-	-	-
		FLO11	4104	-	-	-	38.7	6.6	-	-
	S288C (long)									
		FLO1	4615	18.0	9.4	–	–	–	–	–
		FLO5	3228	8.0	9.6	–	–	–	–	–
		FLO9	3978	13.0	8.3	–	–	–	–	–
		FLO10	3508	–	3.8	4.4	–	–	–	–
		FLO11	4104	–	–	–	38.7	6.6	–	–
	CBS 1483									
		FLO1 (ScVI)	1038	–	–	–	–	–	–	–
		FLO5 (ScI)	2603	1	11.1	–	–	–	–	–
		FLO9 (ScI)	2967	5	15.4	–	–	–	–	–
		FLO11 (ScIX)	2787	–	–	–	14.1	5.6	–	–
S. eubayanus										
	CBS 12357									
		FLO1	5517	–	–	–	–	–	24.7	2.8
		FLO5	1325	–	–	–	–	–	–	–
		FLO9 (SeI)	4752	–	–	–	–	–	8.3	45.9
		FLO9 (SeVI)	3480	–	–	–	–	–	–	–
		FLO9 (SeX)	4041	–	–	–	–	–	7.4	20.1
		FLO9 (SeXII)	3321	–	–	–	–	–	–	10.2
		FLO10 (SeXI)	4128	–	–	–	–	–	–	–
		FLO11 (SeIX)	4149	–	–	–	–	–	–	–
	CBS 1483									
		FLO5 (SeI)	1945	–	–	–	–	–	4.9	2.8
		FLO5 (SeI)	391	–	–	–	–	–	–	–
		FLO5 (SeVI)	3765	–	–	–	–	–	–	–
		FLO5 (SeXI)	2582	–	–	–	–	–	4.9	2.8
		FLO9 (SeI)	2100	–	–	–	–	–	3.0	3.8
		FLO9 (SeXII)	2892	–	–	–	–	–	–	6.3
		FLO10 (SeVI)	3378	–	–	–	–	–	–	–
		FLO11 (SeIX)	3909	–	–	–	–	–	–	–

ment of previously published partial and complete LgFLO1 sequences did not confirm the presence of a similar ORF in CBS 1483. Moreover, the annotated FLO genes had higher identity with *S. eubayanus* and *S. cerevisiae* FLO genes, than with LgFLO1. Therefore, flocculation is likely to rely on one or several of the identified FLO genes from *S. cerevisiae* or *S. eubayanus* subgenomes (Table 4.3).

4.3.3 Sequence heterogeneity in CBS 1483

As other Group 2 *S. pastorianus* strains, CBS 1483 displays heterozygosity between different copies of its *S. cerevisiae* subgenome [208]. We therefore systematically identified heterozygous nucleotides in its genome and investigated the ORFs with allelic variation. Using 159x coverage of paired-end Illumina library of CBS 1483, we found a total of 6,367 heterozygous SNPs across the genome (Additional File 1F in [212]). Although the heterozygous SNPs are present across the whole genome, they affect primarily the S. cerevisiae sub-genome, with the majority clustered around centromeres (Figure 4.2). Of these

positions, 58% were located within ORFs, resulting in 896 ORFs with allelic variation consisting of 1 to 30 heterozygous nucleotides. A total of 685 ORFs showed heterozygosity which would result in amino acid sequence changes, including 16 premature stop codons, 4 lost stop codons and 1566 amino acid substitutions (Additional File 1F [212]). Gene ontology analysis of the ORFs affected by heterozygous calls revealed no significant enrichment in processes, functions of compartments. However, it should be noted that several industrially-relevant genes encoded more than one protein version, such as: the BDH1 and BDH2 genes, encoding butane-diol dehydrogenases involved in reduction of the off flavour compound diacetyl [299], the FLO5 and FLO9 genes encoding flocculins [298], and the OAF1 gene encoding a regulator of ethyl-ester production pathway [300].

4.3.4 Structural heterogeneity in CBS 1483 chromosomes

We investigated whether information about structural heterogeneity between chromosome copies could be recovered despite the fact that current assembly algorithms reduce genome assemblies to consensus sequences. Information about structural and sequence variation between different chromosome haplotypes is not captured by consensus assemblies. However, raw read data contains information for each chromosome copy. To identify structural heterogeneity, we identified ORFs whose predicted copy number deviated from that of the surrounding region in the chromosome based on read coverage analysis (Figure S3 in [212]). We found 213 ORFs with deviating copy number (Additional File 1G in [212]). While no enrichment was found by gene ontology analysis, many of these ORFs are located in subtelomeric regions [225]. Nevertheless, a few regions contained adjacent ORFs with deviating copy number, indicating larger structural variation between chromosome copies. For example, 21 consecutive ORFs in the right-end of the ScXV appear to have been deleted in 2 of the 3 chromosome copies (Figure S3 in [212]). UIP3, one of the genes with deviating copy number, was located on the right arm of chromosome ScI. This region was previously identified as having an additional copy in CBS 1483, although it could not be localized based on short read data [137]. The assembly graph showed two possible structures for ScI, which were collapsed into a single contig in the final assembly (Figure 4.1A). Sequence alignment, gene annotations and sequencing coverage indicated two versions of the ScI contigs: one with and one without the gene UIP3 (Figure 4.1B). Sequence alignments of raw-long-reads revealed five reads (from 20.6 to 36.7 Kbp) linking the right arm of ScI to the left arm of ScXIV at position ~561 Kbp (Figure 4.1C). This location corresponded to a Ty-2 repetitive element; known to mediate recombination within *Saccharomyces* genomes [224]. In addition to the increased coverage of the right arm of ScI, the left arm of ScXIV showed decreased sequencing coverage up until the ~561 Kbp position. Together, these results suggest that the left arm of one copy of ScXIV was replaced with an additional copy of the right arm of ScI (Figure 4.1D). As no reads covered both the recombination locus and the UIP3 locus, it remained unclear if UIP3 is present in the ScI copy translocated to chromosome ScXIV. The resolution of two alternative chromosome architectures of ScI and ScXIV illustrates the ability of long-read alignment to resolve structural heterozygosity.

4.3.5 Differences between Group 1 and 2 genomes do not result from separate ancestry

S. pastorianus strains can be subdivided into two separate groups—termed Group 1 and Group 2—based on both phenotypic [301] and genomic features [208, 269]. However, the ancestral origin of each group remains unclear. The two groups may have emerged by independent hybridization events [277]. Alternatively, Group 1 and Group 2 strains may originate from the same hybridization event, but Group 2 strains later hybridized with a different *S. cerevisiae* strain [208]. In both cases, analysis of the provenance of genomic material from Group 1 and Group 2 genomes could confirm the existence of separate hybridization events if different ancestries are identified. Pan-genomic analysis of *S. cerevisiae* strains indicated that their evolution was largely non-linear, involving frequent horizontal gene transfer and sexual backcrossing events [73]. Especially if the evolutionary ancestry of *S. pastorianus* involves admixture of different *S. cerevisiae* genomes [208], approaches considering only linear evolution such as phylogenetic trees are insufficient [302]. Complex, non-linear evolutionary relationships could be addressed with network approaches [303]. However, such algorithms are not yet fully mature and would involve extreme computational challenges [174, 304].

Therefore, we developed *Alpaca*: a simple and computationally inexpensive method to investigate complex non-linear ancestry via comparison of sequencing datasets [213]. *Alpaca* is based on short-read alignment of a collection of strains to a partitioned reference genome, in which the similarity of each partition to the collection of strains is independently computed using kmer sets [213]. Reducing the alignments in each partition to kmer sets prior to similarity analysis is computationally inexpensive. Phylogenetic relationships are also not recalculated, but simply inferred from previously available information on the population structure of the collection of strains [213]. The partitioning of the reference genome enables the identification of strains with high similarity to different regions of the genome, enabling the identification of ancestry resulting from non-linear evolution. Moreover, since similarity analysis is based on read data, heterozygosity is taken into account.

We used *Alpaca* to identify the most similar lineages for all non-overlapping 2 Kbp sub-regions in the genome of the Group 2 *S. pastorianus* strain CBS 1483 using the reference dataset of 157 *S. cerevisiae* [15] strains and 29 *S. eubayanus* [265]. We inferred population structures for both reference datasets by using previously defined lineages of each strain along with hierarchical clustering based on genome similarity using *MASH* [177]. For the *S. eubayanus* subgenome, almost all sub-regions of CBS 1483 were most similar to strains from the Patagonia B—Holartic lineage [265] (Figure 4.3). In fact, 68% of all sub-regions were most similar to the Tibetan isolate CDFM21L.1 (65) and 27% to two highly-related North-American isolates (Figure S4 in [212]), indicating a monophyletic ancestry of the *S. eubayanus* genome. Analysis of *S. pastorianus* strains CBS 2156 and WS 34/70 (Group2), and of CBS 1503, CBS 1513 and CBS 1538 (Group 1), indicated identical ancestry of their *S. eubayanus* subgenomes (Figure S4 in [212]). Overall, we did not discern differences in the *S. eubayanus* subgenomes of *S. pastorianus* strains, which seem to descend from a strain of the Patagonia B—Holartic lineage and which is most closely related to the Himalayen isolate CDFM21L.1.

In contrast, for the *S. cerevisiae* sub-genome of CBS 1483, the most similar *S. cerevisiae*

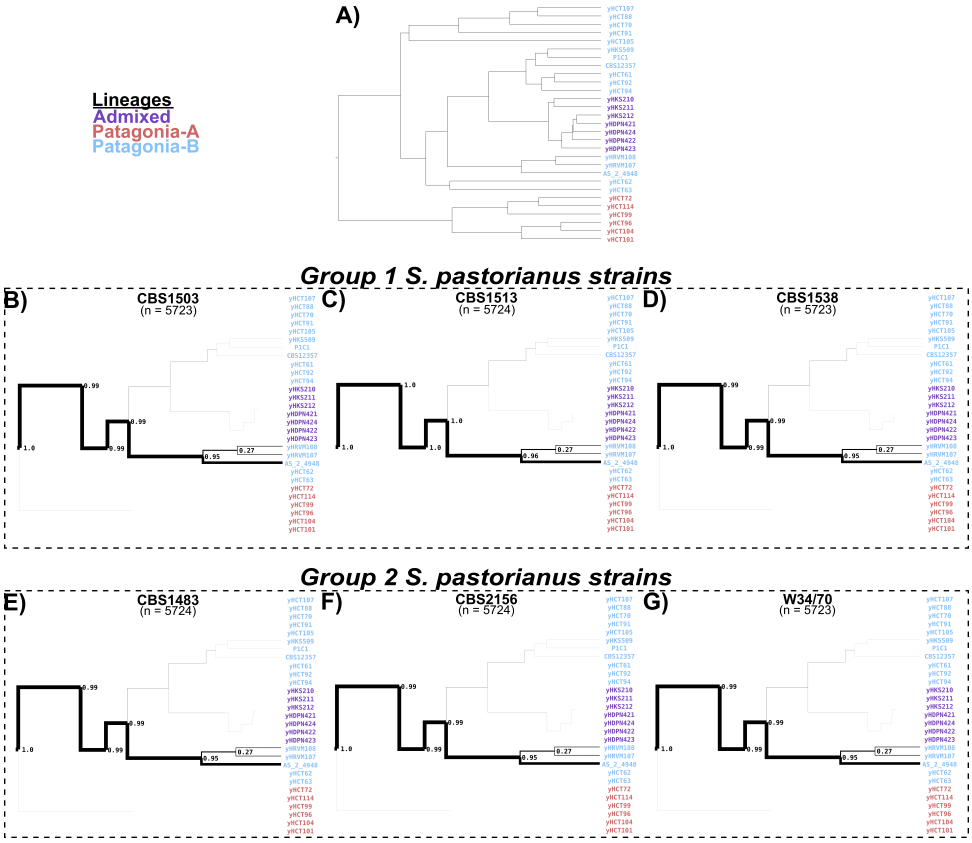


Figure 4.3: Tree-tracing of the genome-scale similarity across the *S. eubayanus* (sub-)genomes of Group 1 and 2 *S. pastorianus* strains, as determined using *Alpaca*. The frequency at which a genome from the reference data set of 29 *S. eubayanus* genomes from Peris *et al* [265] was identified as most similar for a sub-region of the CBS 1483 genome is depicted. The reference dataset is represented as a population tree, upon which only lineages with similarity are indicated with a thickness proportional to the frequency at which they were found as most similar ('N' being the total sum of the number of times all samples appeared as top-scoring). The complete reference population tree (A), the genomes of Group 1 strains CBS 1503, CBS 1513 and CBS 1538 (B-D) and for the genomes of Group 2 strains CBS 1483, CBS 2156 and WS34/70 (E-G) are shown.

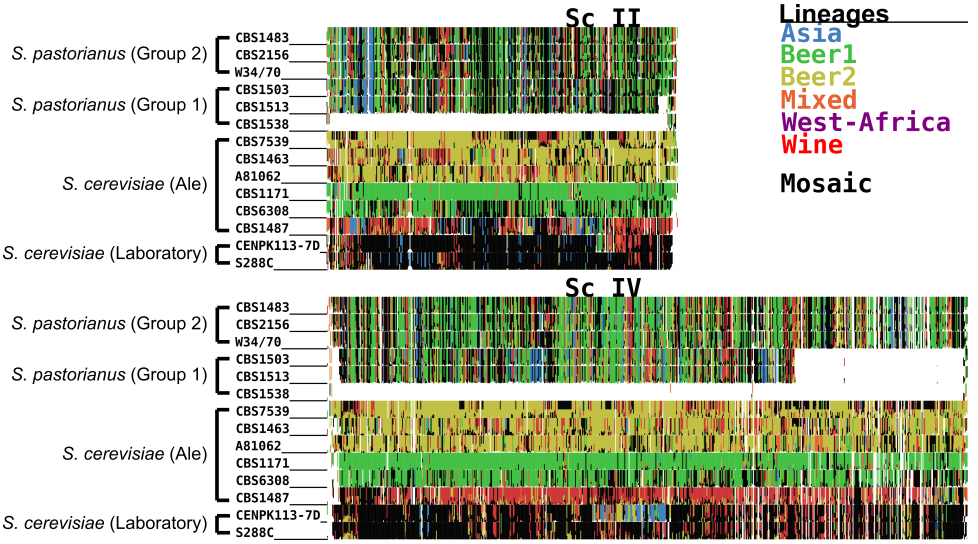


Figure 4.4: Similarity profiles of the *S. cerevisiae* (sub-)genomes of various *Saccharomyces* strains, as determined using *Alpaca* for chromosomes Sc II and IV. Each *S. cerevisiae* chromosome of the CBS 1483 assembly was partitioned in non-overlapping sub-regions of 2 Kbp. The colors represent the most similar lineages based on kmer similarity of 157 *S. cerevisiae* strains from Gallone *et al.* [15]: Asia (blue), Beer1 (green), Beer2, (gold), Mixed (orange), West-Africa (purple), Wine (red). Mosaic strains are shown in black and ambiguous or low-similarity sub-regions in white. Similarity patterns are shown for the Group 2 *S. pastorianus* strains CBS 1483, CBS 2156, WS34/70 and Hei-A, for the Group 1 *S. pastorianus* strains CBS 1503, CBS 1513 and CBS 1538, for *S. cerevisiae* ale-brewing strains CBS 7539, CBS 1463, A81062, CBS 1171, CBS 6308 and CBS 1483, and for S. cerevisiae laboratory strains CEN.PK113-7D and S288C. Similarity profiles for all chromosomes in the *S. cerevisiae* (sub-)genomes are shown in Figure S5 in [212].

strains varied across the sub-regions of every chromosome (Figure 4.4 and S5 in [212]). No strain of the reference dataset was most similar for more than 5% of sub-regions, suggesting a high degree of admixture (Figure 5 4.4 and S6 [212]). However, 60% of sub-regions were most similar to the Beer 1 lineage, 12% were most similar to the Wine lineage and 10% to the Beer 2 lineage [15]. In order to determine Alpaca's ability to differentiate genomes with different admixed ancestries, we analysed the genomes of 8 *S. cerevisiae* strains: six ale-brewing strains and the laboratory strains CEN.PK113-7D and S288C. The strains CBS 7539, CBS 1463 and A81062 were identified as similar to the Beer 2 lineage, CBS 1171 and CBS 6308 as similar to the Beer 1 lineage, CBS 1487 as similar to the Wine lineage, and CEN.PK113-7D and S288C as similar to the mosaic laboratory strains (Figure 4.4 and S5 in [212]). In addition, the distribution of similarity over the *S. cerevisiae* population tree differed per strain (Figure 4.5 and S6 in [212]). While no single strain was most similar for more than 8% of the sub-regions for CBS 1487 and CBS 6308, for CBS 7539 67% of sub-regions were most similar to the strain beer002. As both beer002 and CBS 7539 are annotated as Bulgarian beer yeast [15, 283], this similarity likely reflects common origin. The different similarity profiles of all *S. cerevisiae* strains indicate that Alpaca can differentiate different ancestry by placement of genetic material within the *S. cerevisiae* population tree, whether a genome has a linear monophyletic origin or a non-linear polyphyletic origin.

To identify possible differences in genome compositions within the *S. cerevisiae* subgenomes of *S. pastorianus*, we analysed other Group 1 and 2 strains using *Alpaca*, including an isolate of the Heineken A-yeast® lineage (Hei-A), which was isolated in 1886 and represents one of the earliest pure yeast cultures. Whole genome sequencing, alignment to the CBS 1483 assembly and sequencing coverage analysis revealed that the ploidy of the Hei-A isolate corresponds to that of a Group 2 strain (Figure S7 in [212]). Analysis of Hei-A and the other *S. pastorianus* Group 2 strains CBS 2156 and WS 34/70 using Alpaca yielded almost identical patterns of similarity at the chromosome-level as CBS 1483 (Figure 4.4 and S5 in [212]). Moreover, similarity was distributed across the *S. cerevisiae* population tree almost identically as in CBS 1483 (Figure 4.5 and S6 in [212]). The Group 1 *S. pastorianus* strains CBS 1503, CBS 1513 and CBS 1538 displayed different patterns of similarity at the chromosome-level relative to Group 2 strains. While various chromosome regions harboured almost identical similarity patterns, some regions differed significantly, such as: ScI, the middle of ScIV, the left arm of ScV, ScVIII, the right arm of ScIX, ScX-SeX, ScXI and ScXIII (Figure 4.4 and S5 in [212]). However, at the genome level, similarity was distributed across the *S. cerevisiae* population tree almost identically as in Group 2 strains, except for a slightly higher contribution of the Beer 2 and Wine lineages, at the expense of a lower contribution of the Beer 1 lineage (Figure 4.5 and S6 in [212]). The almost identical distribution of all Group 1 and Group 2 strains over the *S. cerevisiae* population tree indicate that they have the same *S. cerevisiae* ancestry. The spread of similarity across the *S. cerevisiae* population tree advocates for an admixed, possibly heterozygous ancestry of the *S. cerevisiae* subgenome of *S. pastorianus*. Furthermore, the different patterns of similarity at the chromosome level between both groups are compatible with an initially heterozygous *S. cerevisiae* subgenome which was subjected to independent loss of heterozygosity events in each group, resulting in differential retention of each haplotype. The lower relative contribution of Beer 1 strains in Group 1 strains may be explained by the complete absence of *S. cerevisiae* chromosomes with high similarity to Beer1 strains, such as ScV,

ScXI and ScXv-ScXI.

4.4 Discussion

In this study, we used Oxford Nanopore Technology's (ONT) MinION sequencing platform to study the genome of CBS 1483, an alloaneuploid Group 2 *S. pastorianus* strain. The presence of extensively aneuploid *S. cerevisiae* and *S. eubayanus* subgenomes substantially complicates analysis of *S. pastorianus* genomes [137]. We therefore explored the ability of ONT sequencing to generate a reference genome in the presence of multiple non-identical chromosome copies, and investigated the extent to which structural and sequence heterogeneity can be reconstructed. Despite its aneuploidy, we obtained a chromosome-level genome haploid assembly of CBS 1483 in which 29 of the 31 chromosomes were assembled in a single contig. Comparably to assemblies of euploid *Saccharomyces* genomes [207, 234, 235, 237, 260, 271], ONT sequencing resulted in far lesser fragmentation and in the addition of considerable sequences compared to a short-read based assembly of CBS 1483, notably in the subtelomeric regions [137]. The added sequences enabled more complete identification of industrially-relevant subtelomeric genes such as the MAL genes, responsible for maltose and maltotriose utilisation [285–287], and the FLO genes, responsible for flocculation [231, 290, 291]. Due to the instability of subtelomeric regions [223, 225, 233], the lack of reference-based biases introduced by scaffolding allows more certainty about chromosome structure [207]. Since subtelomeric genes encode various industrially-relevant traits [228–231], their mapping enables further progress in strain improvement of lager brewing yeasts. Combined with recently developed Cas9 gene editing tools for S. pastorianus [209], accurate localisation and sequence information about subtelomeric genes is critical to investigate their contribution to brewing phenotypes by enabling functional characterization [305].

Despite the presence of non-identical chromosome copies in CBS 1483, the genome assembly only contained one contig per chromosome. While the assembly did not capture information about heterogeneity, mapping of short-read data enabled identification of sequence heterozygosity across the whole genome. In previous work, two alternative chromosome structures could be resolved within a population of euploid *S. cerevisiae* strain CEN.PK113-7D by alignment of long-reads [207]. Therefore, we evaluated the ability to identify structural heterogeneity by aligning long-read data to the assembly. Indeed, long-read alignments enabled the identification of two versions of chromosome ScI: with and without an internal deletion of the gene UIP3. Furthermore, the length of long-reads enabled them to span a TY-element, revealing that one of the copies of the right arm of ScI was translocated to the left arm of ScXIV. While the two alternative structures of ScI constitute a first step towards the generation of chromosome copy haplotypes, long-reads only enabled the hypothesis-based resolution of suspected heterogeneity. Assembly algorithms which do not generate a single consensus sequence per chromosome are emerging [170, 306]. However, haplotyping is particularly difficult in aneuploid and polyploid genomes due to copy number differences between chromosomes [306]. A further reduction of the relatively high error rate of long-reads, or the use of more accurate long-read sequencing technologies, could simplify the generation of haplotype-level genome assemblies in the future by reducing noise [307].

We used the chromosome-level assembly of CBS 1483 to study the ancestry of *S. pas-*

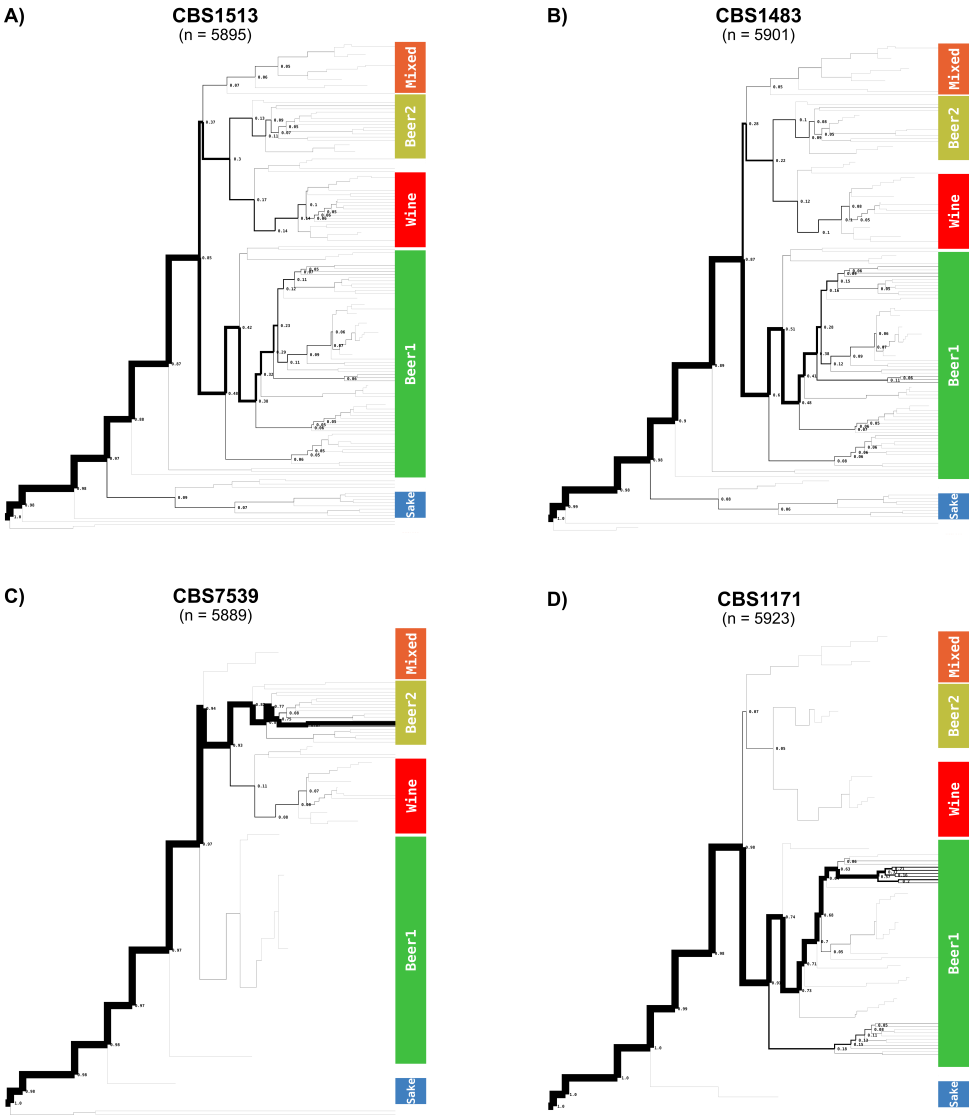


Figure 4.5: Tree-tracing of the genome-scale similarity across the *S. cerevisiae* (sub-)genomes of various *Saccharomyces* strains, as determined using *Alpaca*. The frequency at which a genome from the reference data set of 157 *S. cerevisiae* strains from Gallone *et al.* [15] was identified as most similar for a sub-region of the CBS 1483 genome is depicted. The reference dataset is represented as a population tree, upon which only lineages with similarity are indicated with a thickness proportional to the frequency at which they were found as most similar ('N' being the total sum of the number of times all samples appeared as top-scoring). The genomes of *S. pastorianus* Group 1 strain CBS 1513 (A), of *S. pastorianus* Group 2 strain CBS 1483 (B), of *S. cerevisiae* strain CBS 7539 and of *S. cerevisiae* strain CBS 1171 are shown. The tree-tracing figures of *S. pastorianus* Group 1 strains CBS 1503 and CBS 1538, of *S. pastorianus* Group 2 strains CBS 2156, WS34/70 and Hei-A, and of *S. cerevisiae* strains CBS 1463, A81062, CBS 6308, CBS 1487, CEN.PK113-7D and S288C are shown in Figure S6 in [212].

torianus genomes. Due to the importance of non-linear evolution in the domestication process of *Saccharomyces* strains [73], and to the admixed hybrid nature of *S. pastorianus* [208, 265], we used the newly-developed method *Alpaca* to analyse the ancestry of CBS 1483 instead of classical phylogenetic approaches using reference datasets of S. cerevisiae and S. eubayanus strains [15, 265]. All *S. pastorianus* genomes displayed identical distribution of similarity across the reference *S. eubayanus* population tree, both at the chromosome and whole-genome level. All *S. pastorianus* genomes also showed identical distribution of similarity across the reference *S. cerevisiae* population tree at the whole genome level; however, Group 1 and Group 2 strains displayed different similarity patterns at the chromosome level. The absence of differences in the *S. cerevisiae* genome at the whole genome level and recurrence of identical chromosomal break points between Group 1 and 2 strains discredit previous hypotheses of different independent hybridization events in the evolution of Group 1 and 2 strains [208, 277]. Instead, these results are compatible with the emergence of Group 1 and 2 strains from a single shared hybridization event between a homozygous *S. eubayanus* genome closely related to the Tibetan isolate CDFM21L.1 and an admixed heterozygous *S. cerevisiae* genome with a complex polyphyletic ancestry. Loss of heterozygosity is frequently observed in *Saccharomyces* genomes [73, 308], and therefore likely to have affected both the genomes of Group 1 and 2 strains [208, 309, 310]. The different chromosome-level similarity patterns in both groups likely emerged through different loss of heterozygosity events in Group 1 and 2 strains [309, 310]. In addition, the lower *S. cerevisiae* chromosome content of Group 1 is consistent with observed loss of genetic material from the least adapted parent during laboratory evolution of *Saccharomyces* hybrids [311–315]. In this context, the lower *S. cerevisiae* genome content of Group 1 strains may have resulted from a rare and serendipitous event. For example, chromosome loss has been observed due to unequal chromosome distribution from a sporulation event of a allopolyploid *Saccharomyces* strain [315]. Such mutant may have been successful if loss of S. cerevisiae chromosomes provided a selective advantage in the low-temperature lager brewing environment [311, 311]. The loss of the *S. cerevisiae* subgenome may have affected only Group 1 strains due to different brewing conditions during their domestication. However, the high conservation of similarity within Group 1 and Group 2 strains indicate that the strains within each Group are closely related, indicating a strong population bottleneck in their evolutionary history.

Such a bottleneck could have been caused by the isolation and propagation of a limited number *S. pastorianus* strains, which may have eventually resulted in the extinction of other lineages. The first *S. pastorianus* strains isolated in 1883 by Hansen at the Carlsberg brewery were all Group 1 strains [210, 316]. Due to the industry practice of adopting brewing methods and brewing strains from successful breweries, Hansen's Group 1 isolates likely spread to other breweries as these adopted pure culture brewing [2]. Many strains which were identified as Group 2 by whole genome sequencing were isolated in the Netherlands [208, 269]: Elion isolated the Heineken A-yeast® in 1886 [317], CBS 1484 was isolated in 1925 from the Oranjeboom brewery [269], CBS 1483 was isolated in 1927 in a Heineken brewery [137], and CBS 1260, CBS 2156 and CBS 5832 were isolated from unknown breweries in the Netherlands in 1937, 1955 and 1968, respectively [269, 318]. Analogously to the spread of Group 1 strains from Hansen's isolate, Group 2 strains may have spread from Elion's isolate. Both Heineken and Carlsberg distributed their pure cul-

ture yeast biomass to breweries over Europe and might therefore have functioned as an evolutionary bottleneck by supplanting other lineages with their isolates [319, 320]. Overall, our results support that the differences between Group 1 and 2 strains emerged by differential evolution after an initial shared hybridization event, and not by a different *S. eubayanus* and/or *S. cerevisiae* ancestry.

Beyond its application in this study, we introduced Alpaca as a method to evaluate non-linear evolutionary ancestry. The use of short-read alignments enables Alpaca to account for sequence heterozygosity when assessing similarity between two genomes and is computationally inexpensive as they are reduced to kmer sets. Moreover, *Alpaca* leverages previously determined phylogenetic relationships within the reference dataset of strains to infer the evolutionary relationship of the reference genome to the dataset of strains. Due to the presence of non-linear evolutionary processes in a wide range of organisms [321, 322], the applicability of Alpaca extends far beyond the *Saccharomyces* genera. For example, genetic introgressions from *Homo neanderthalensis* constitute about 1% of the human genome [323]. Horizontal gene transfer is even relevant across different domains of life: more than 20% of ORFs of the extremely thermophilic bacteria *Thermotoga maritima* were more closely related to genomes of Archaea than to genomes of other Bacteria [324]. Critically, horizontal gene transfer, backcrossing and hybridization have not only played a prominent role in the domestication of Saccharomyces yeasts [73], but also in other domesticated species such as cows, pigs, wheat and citrus fruits [325–328]. Overall, Alpaca can significantly simplify the analysis of new genomes in a broad range of contexts when reference phylogenies are already available.

4.5 Conclusion

With 29 of the 31 chromosomes assembled in single contigs and 323 previously unassembled genes, the genome assembly of CBS 1483 presents the first chromosome-level assembly of a *S. pastorianus* strain specifically, and of an alloaneuploid genome in general. While the assembly only consisted of consensus sequences of all copies of each chromosome, sequence and structural heterozygosity could be recovered by alignment of short and long-reads to the assembly, respectively. We developed *Alpaca* to investigate the ancestry of Group 1 and Group 2 *S. pastorianus* strains by computing similarity between short-read data from *S. pastorianus* strains relative to large datasets of *S. cerevisiae* and *S. eubayanus* strains. In contrast with the hypothesis of separate hybridization events, Group 1 and 2 strains shared similarity with the same reference *S. cerevisiae* and *S. eubayanus* strains, indicating shared ancestry. Instead, differences between Group 1 and Group 2 strains could be attributed to different patterns of loss of heterozygosity subsequent to a shared hybridization event between a homozygous *S. eubayanus* genome closely related to the Tibetan isolate CDFM21L.1 and an admixed heterozygous *S. cerevisiae* genome with a complex polyphyletic ancestry. We identified the Heineken A-yeast® isolate as a Group 2 strain. We hypothesize that the large differences between Group 1 and Group 2 strains and the high similarity within Group 1 and 2 strains result from a strong population bottleneck which occurred during the isolation of the first Group 1 and Group 2 strains, from which all currently known *S. pastorianus* strains descend. Beyond its application in this study, the ability of Alpaca to reveal non-linear ancestry without requiring heavy computations presents a promising alternative to phylogenetic network analysis to investigate

horizontal gene transfer, backcrossing and hybridization.

5

A streaming algorithm to infer species composition in *Saccharomyces* hybrid genomes

5.1 Introduction

Saccharomyces yeasts are central organisms in various industrial applications. Historically, humans have used yeast for agricultural purposes, such as bread making and alcohol fermentation [329–331]. In the past few decades, *Saccharomyces* yeasts have been adapted for biotechnological purposes, such as the production of therapeutics and alternative energy sources [72, 332, 333]. We now know that yeasts used in industry typically fall under the *Saccharomyces sensu strictu* genus made up of eight different species: *S. arboricola*, *S. cerevisiae*, *S. eubayanus*, *S. jurei*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, and *S. uvarum* [212, 256, 260, 263, 334–338]. The most notable is *S. cerevisiae*: a widely adopted model organism actively used in various industrial processes. But other species—such as *S. eubayanus*, *S. kudriavzevii*, and *S. uvarum*—are also actively used in wine, cider, and beer fermentation [212, 256, 263, 335, 336, 339, 340].

Whole-genome sequencing (WGS) is providing extensive insights in the genomic diversity of the *Saccharomyces sensu strictu* genus. For example, a recent study of more than 1,000 global *S. cerevisiae* isolates has provided in-depth look into it's evolutionary history, specifically, a likely "Out-of-China" evolutionary origin [73]. WGS projects sequencing tens to hundreds of *S. uvarum* and *S. eubayanus* global isolates are similarly providing insights in their respective evolutionary histories. [264, 265]. The latter species being only recently isolated from Patagonia, Argentina, and proposed as an individual member of the *Saccharomyces sensu strictu* genus in 2011 [262]. Similarly, *S. jurei* was also only recently isolated in high altitudes in Southern France and proposed as a new species and member in 2017 [336].

Additionally, WGS of *Saccharomyces sensu strictu* isolates have highlighted their non-linear evolution, such as alcohol-fermenting-yeasts which can be (natural) hybrids of two or more *Saccharomyces* species [212, 256, 263, 335]. Lager-beers are brewed with *S. pasto-*

rianus strains—a natural hybrid harbouring the genomes of *S. cerevisiae* and *S. eubayanus* in the same nucleus [212, 339]. Some ciders are brewed with *S. bayanus*—a natural triple hybrid between *S. cerevisiae*, *S. eubayanus*, and *S. uvarum* [256, 341]. Similarly, some wines are brewed with hybrids from different combinations of *Saccharomyces* species [340, 342, 343]. In all cases, the genomic contribution from each of the corresponding parental species varies, highlighting complex evolutionary histories and environmental adaptations [208, 212, 263]. Furthermore, hybridization is not only reserved to combinations of multiple species: many brewing strains are products of (artificial) crossing of different strains from the same species, leading to mosaic genomes and admixed populations [15].

With the affordability of WGS, brewing industries and research institutions alike will (or already are) sequencing large collections of yeast species, many of which could include novel yeast strains and hybrids [210, 212, 263]. For example, the Westerdijk Fungal Biodiversity Institute (CBS-KNAW) houses the world largest collection of fungal specimens and is known to house various *Saccharomyces* strains and hybrids. Determining the ancestral origin and species composition of such isolates and *Saccharomyces* metagenomic communities is thus an immediate interest to the yeast reasearch community [208, 212, 263, 336, 344].

Various methods are used for determining the species composition and ancestral origins of *Saccharomyces* yeasts. Focusing on WGS, (whole-)genome comparisons provide robust evolutionary analysis, either by reporting genome-similarity from whole-genome alignments, or constructing phylogenetic trees from single-nucleotide variants of one or more cores genes [137, 208, 262–265]. The recent feasibility of constructing near-complete genome assemblies from long-read sequencing is also enabling in-depth look at complex evolutionary histories due intra/inter-strain hybridization, as we previously reported [212, 213]. Although these approaches can provide high-resolution insights in the evolutionary histories of yeast species, industrial yeasts are often aneuploid—that is, possess a variable number of copies per chromosome—with heterozygous variation [137, 208, 212]. *De novo* genome assemblies of such isolates can thus lead to fragmented assemblies and/or consensus representations of one or more haplotypes, challenging traditional evolutionary inference methods that assume a haploid-configuration [137, 208, 212].

An alternative approach for determining the species composition and ancestral origins of *Saccharomyces* yeasts is via read-alignments to a reference genome. Coverage information using a concatenated reference of more than one species can help assess the hybrid-nature of an isolate [208, 263, 264], while a phylogenetic tree from single-nucleotide variants can be used to infer its evolutionary origins [15, 73, 264, 265]. Since the former approach leads to multi-mapping reads due to sequence homology between *Saccharomyces* species, read-alignments must be curated to account only pre-defined unique-gene markers, and/or filtered via some criteria such as sequence-divergence [263, 264]. Additionally, the aneuploid nature of industrial yeasts complicates phylogenetic-tree constructions as it is not entirely clear how to account for sequence heterozygosity [345–349].

In either approach, the choice of a reference(s) is critical to properly assess the species composition of a *Saccharomyces* hybrid genome, which circularly requires prior knowledge of the evolutionary history of the genome. A method that can automatically assess the global *Saccharomyces* species composition in a sequencing dataset without any

prior known can circumvent computational challenges in analysing hybrid *Saccharomyces* yeasts and facilitate in-depth downstream evolutionary analysis.

We thus developed *Redwood2*, a "ready-to-run" alignment-free method to quickly assess the species composition of an input WGS-dataset for a *Saccharomyces* isolate without any pre-processing of the data. Our method is largely inspired by recent work on probabilistic representations of genomes, and resembles the computational problem of classifying taxa in a metagenomic dataset. In short, we can represent a phylogenetic tree of hundreds of available sequencing datasets from all eight species in the *Saccharomyces sensu strictu* by utilizing genome sketches, sequence-bloom trees, and the HyperLogLog algorithm. By adapting a k-mer based streaming-algorithm, we can quickly calculate significant containments of one or more nodes (e.g. species, lineages, and strains) from the tree in given a WGS-dataset—all while considering the evolutionary relationships described in the phylogenetic tree.

We first generalize the problem of inferring the species composition of a hybrid genome as the intersection of multiple sets of k-mer sequences, and extend it to account for hybridizations of intra-/inter-species and lineage members. We then benchmark our method on computationally simulated assemblies of single and hybrid *Saccharomyces* genomes showcasing the sensitivity and accuracy of our method. Finally, we applied our method on available WGS-datasets of recently reported hybrid *Saccharomyces* isolates and validate their reported evolutionary histories and hybridization events.

5.2 Methods

5.2.1 The set-containment problem in the context of possible hybridization events from a phylogenetic tree

Consider a microbial genome represented as a set of k-mers, regarded as Q. Now consider the k-mer sets of two additional microbial samples, A and B, where $A \cap B = \emptyset$. Under the scenario in which Q is a hybrid genome from a direct hybridization of A and B, then A and B are subsets of Q (see Figure 5.1A). The genomic contribution of A and B in Q is thus $\frac{|A \cap Q|}{|Q|} = 0.5$ and $\frac{|B \cap Q|}{|Q|} = 0.5$. Under the scenario of only a partial hybridization, one would observe *fractional genomic contributions* (FGC): for example, a hybridization event from 100% of genome A and 50% of genome B, will yields FGC's of $\frac{|A \cap Q|}{|Q|} = 0.667$ and $\frac{|B \cap Q|}{|Q|} = 0.333$.

The above calculations assume that the genomes of A and B share no sequence similarity, which is pragmatically invalid when considering genomes from similar evolutionary histories; for example, if A and B share a common ancestor (see Figure 5.1B). Therefore, if A but not B is contained in Q, one would still expect to observe $|A \cap B|$ k-mers if testing whether B is contained in Q. Thus, the FGC's of A and B in Q needs to be corrected for shared k-mers, for example, the number of unique k-mers in A found in Q but not in B divided by the number of unique k-mers found in Q, or $\frac{|A \cap Q| - |A \cap B|}{|Q \cap (A \cup B)|}$.

To generalize in the context of a tree, let $FGC(N) = \frac{\Theta_{U_N}}{\Theta_{R_N}}$, where Θ_{U_N} is the number of unique k-mers uniquely contained in node N or any of its children, and Θ_{R_N} is the

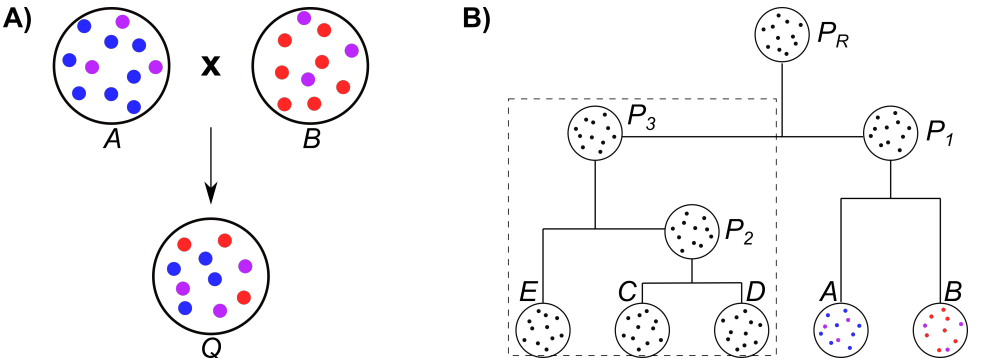


Figure 5.1: Genome sketches in hybrid-genomes and phylogenetic trees. (A) Sketch representations of genomes A and B, where blue dots represent k-mers unique to A, red dots indicates k-mers unique to B, and purple dots represent k-mers shared by both genomes. A genome-hybridization event would yield genome Q with the illustrated sketch. (B) Phylogenetic tree containing genomes A and B along with genomes C, D, and E. Nodes P_1, P_2, P_3, and P_R are parent nodes. The dashed square highlights a sub-tree with P_3 as the root, used as a running example in 5.2.1 to illustrate *Redwood2*'s calculation for estimating the species composition of a node in a phylogenetic tree.

total possible number of unique k-mers contained in N after adjusting for shared k-mers with other nodes that are not part of the sub-tree with N as the root. Additionally, $\Theta_{R_N} = \Theta_{T_N} - \Theta_{S_N}$, where Θ_{T_N} is the total number of unique k-mers in a query genome (e.g. Q) that are also in the tree, and Θ_{S_N} is the number unique k-mers that are shared with nodes outside of the sub-tree of N. Note that this implies that there is a sensitivity level: if the A and B are nearly identical, then one cannot confidently differentiate whether the containment observed in Q derives from one or both genomes. However, one can still assess the contribution of the parent node, P_1, representing $(A \cup B)$, in the containment in Q; in this case, $\frac{|P \cap Q|}{|P|}$ since P_1 is the root node (see Figure 5.1B).

Now consider the sub-tree in Figure 5.1B with node P_3 as the root (sub-tree highlighted by the dashed rectangle). Although testing the containment of one of the leafs in query genome, Q, requires one to consider the other additional nodes, the FGC calculations remain the same. For example, testing the containment of node E in Q still requires one to account for the number of k-mers uniquely contained in E and number of k-mers that are shared with nodes outside the tree, such as P_3, C and D. Similarly, the probability used when testing for the significance of the fractional genomic contribution only needs to be adjusted to account for the evolutionary relationships in the other side of the tree. In the running example, when testing for the significance for the fractional genomic contribution from E, then r is $\frac{|E \cap C \cap D|}{|AE \cup C \cup D|}$. This calculation can recursively be computed across any node in a binary tree, regardless of the size 5.1B.

Statistical significance in observing x-number of k-mers uniquely contained between any node in C can be tested via a Binomial distribution [177, 350]. First, to test whether any node in the tree is significantly contained in Q, we can compute $Prob(x_T, n_T, r_T)$ via a binomial test, where n_T is the total number of possible k-mers in the tree, r_T is the probability of a k-mer being contained in the tree, and x_T is the number of k-mers in Q

observed in the tree [177, 350]. If this is significant, then we can test for the significance of the fractional genomic contribution of a node in the tree in Q via the same calculation, $Prob(x_N, n_N, r_N)$, but here, n_N is the total number of unique k-mers in the tree contained in Q, r_N is the probability of a k-mer being shared with any other node in the tree, and x_N is the number of observed k-mers uniquely contained within testing node N.

Although the calculations for fractional genomic contributions and statistical tests are theoretically possible for large trees, they quickly become computationally infeasible: if dealing with *Saccharomyces* genomes with a minimum genome-length of ~12 Mbp, a single leaf-node will require at least 96 megabytes of memory (assuming k-mers are represented as 64-bit integers and no sequencing errors exist). If the k-mer sets of the leafs and parent-nodes are jointly loaded into memory, the calculations will require more than one gigabyte of memory for a tree of only 6 samples; requiring specialized computational infrastructure to make use of the genomic diversity in published WGS-datasets of hundreds of *Saccharomyces* samples.

In the next section, we describe how the fractional genomic contribution and statistical test calculations can be approximated with high sensitivity and accuracy by recent probabilistic genome-representations.This approximation is implemented in a stand-alone method called, *Redwood2*.

5.2.2 Approximate fractional genome contribution calculations with *Redwood2*

The calculations from the previous section can be approximated by adopting existing work on probabilistic genome representations, which has extensively been discussed in a recent review [107]. Ondov *et al.* [177] previously described the sketch-representation of a genome: a (reduced) k-mer set enabling rapid approximate genome similarity and containment calculations with reduced computational requirements. Other work has followed utilizing sketch-representations: Brown *et al.* [351] combine genome-sketches and sequence bloom trees (a tree storing a bloom-filter per node which themselves are probabilistic sets with reduced computational requirements) to compute taxonomic classifications of genomes. Similarly, Breitwieser *et al.* [352] employ a HyperLogLog algorithm (a probabilistic algorithm estimating the number of unique elements in a set with small memory footprint) to similarly compute taxonomic classifications from pre-classified reads in WGS-datasets. Bradley *et al.* [178] utilize adaptations of all these representations to efficiently search large collections of microbial genome sequences. In follow-up study, Ondov *et al.* [350] proposed a streaming algorithm to compute significant containments of single genomes within metagenomes using sketch-representations. Similarly, we can adapt all these ideas (e.g. genome sketches, sequence bloom-trees, and the HyperLogLog algorithm) to efficiently compute the fractional genomic containment and statistical test calculations for species, lineages, and strains as discussed in the previous section.

The first challenge is to represent a large phylogenetic tree as a data structure that enables efficient calculations of the fractional genomic contribution and statistical tests. In essence, we can utilize a sequence-bloom tree such that each leaf is the k-mer set of a genome stored as a bloom-filter with a specified false-positive rate, f, and a maximum number of inserted elements, i. Each node is thus the union the bloom-filters of their children. Much like in Brown *et al.* and Breitwieser *et al.*, we can query k-mers in a given

genome (or WGS-dataset) across the tree to calculate taxonomic classifications based on the number of k-mers contained in each node [351, 352]. For scalability purposes, we can instead store an s-size bottom-sketch in each leaf, leading to a sequence bloom-tree with significantly reduced computational resources, at the expense of classification sensitivity [350–352].

Algorithm 1 Computing values for Θ_{U_N} and Θ_{T_N}.

```
1:  procedure NODE COUNTER(L_i, S, T)   ▷ L_i: Array of bloom-filters for all leafs in tree; S:
    set of read or contig-sequences; T: binary-tree datastructure
2:      L_T ← Union of all L_i
3:      H_T ← Initialize empty set of hashes
4:      for s ∈ S do
5:          for k,∈ s do
6:              h = hash(k)
7:              if h ∈ L_T then
8:                  H.add(h)
9:              end if
10:         end for
11:     end for
12:     n_counts ← Initialize empty counting hash-table
13:     lca_counts ← Initialize empty counting hash-table
14:     for h ∈ H do
15:         L_sub ← sub-array of leaf IDs, where h ∈ L_i
16:         increment nodes in n_counts using L_sub and T
17:         lca_node ← find LCA of L_sub in T
18:         increment lca_counts[lca_node]
19:     end for
20:     return n_counts and lca_counts
21: end procedure
```

Algorithm 1 describes how to compute the necessary values for Θ_{U_N} and Θ_{T_N} using only a partial representation of the sequence bloom-tree mentioned above (represented as T). These two values are used for calculating the fractional genomic composition, FGC, of a node in a phylogenetic tree (see section 5.2.1). Let L_i be the collection of bloom-filters from all leafs in the tree $\{L_i : i \in I\} | I = \{1, 2, 3, \cdots, n\}$. Given the k-mer set for a query genome, H_Q, let H_T be a subset of k-mers, $H_T \subseteq H_U$, such that all k-mers, k_i in H_T, $k_i \in L_i$. By computing H_T (lines 4-11), Algorithm 1 can construct two hash-tables which can then be used to derive Θ_{U_N} and Θ_{T_N}. More specifically, line 15 identifies all leafs that contain a given k-mer, while line 16 performs a depth-first traversal in T and increments the counter in n_counts whenever for all leafs in line 15 or any node parent of the leafs. Line 17 identifies the lowest-common ancestor (LCA) of the leafs in line 15, and line 18 increments the counter in lca_node for each LCA for each k-mer. As such, the hash-table of n_counts represents the total number k-mers contained in some node irrespective if they appear else where in the tree, or in other words, Θ_{T_N}, while lca_counts stores the number of times a node was the LCA for some k-mer. For every node, N in the tree, Θ_{U_N}

can be calculated by aggregating the counts for all children of N using the corresponding values in lca_counts.

Additionally, Θ_{S_N} can be calculated by correcting for the total possible number of unique k-mers appearing only within the subtree of N. Let $hcard(N)$ be the node-cardinality of N (that is, the unique number of k-mers in the node), which can be approximated using the HyperLogLog algorithm [353]. More specifically, compute the cardinality for each leaf using their respective bottom-sketches with the HyperLogLog algorithm, which stores an array of maximum counts observed for some pre-defined array size [353]. Using a recursive depth-first search, estimate node-cardinality for the remaining nodes (e.g. parents) by merging the arrays (e.g. taking the maximum of the corresponding values). Given the estimated node-cardinalities for all nodes in the tree, let N_A and N_B be two nodes (or leafs) that have the same direct parent node, N_P. The fraction of k-mers shared between N_A and N_B can be calculated as, $shared(N_P) = \frac{hcard(N_P)}{hcard(N_A)+hcard(N_A)}$. As such, the probability of a k-mer in N being shared with other nodes outside of the sub-tree with N as the parent node, p_{shared_N} can be computed by summing the product of the shared fractions. For example, using the tree in Figure 5.1B:

$$p_{shared_N} = shared(N_{P_2}) + shared(N_{P_2}) \cdot shared(N_{P_3}) + shared(N_{P_2}) \cdot shared(N_{P_3}) \cdot shared(N_{P_R})$$

Therefore, $FGC(N) = \frac{\Theta_{U_N}}{|H_T| - (\Theta_{T_N} - \Theta_{S_N} + p_{shared_N} \cdot (\Theta_{T_N} - \Theta_{S_N})}$.

Note that to test whether the calculated $FGC(N)$ for node N is significant, we can use p_{shared_N} for the Binomial test (see Methods 5.2.1).

We implemented these approximations in our stand-alone method, *Redwood2 Redwood2* is split into two-modules: (1) partial sequence-bloom tree construction and (2) k-mer streaming from a query genome. The first module enables users to construct a custom-made partial sequence-bloom tree of any given collection of genomes, as opposed to the pre-built trees provided by this paper (see next sub-section). Briefly, the first module is split into three sub-modules: (1.1) bottom-sketching, (1.2) bloom-filter transformation, and (1.3) partial sequence-bloom tree construction. Sub-module 1.1 uses a native implementation of the bottom-sketching algorithm from Ondov *et al.* [350] to enable consistent use of the same 64-bit hash-functions and lexicographic smallest k-mers throughout the rest of the modules. In sub-module 1.2, the sketches are converted to bloom-filters with a specified false positive rate (0.01 by default). Sub-module 1.3 then uses the bottom-sketches and bloom-filters to construct the partial sequence-bloom tree and calculate node-cardinalities using the HyperLogLog algorithm using Algorithm 2—all stored in a single portable binary file. Note that this sub-module expects a binary phylogenetic tree in Newick format to guide the construction of the partial sequence-bloom tree. Alternatively, a pairwise distance matrix of all leafs can be provided to cluster them into a dendrogram which is then used as the tree.

Finally, module 2 in *Redwood2* streams all k-mers in a given query genome to calculate the fractional genomic containments for all nodes in the tree, enabling one to identify significant containments of species, lineages, and strains. This module uses an adaptation of the streaming algorithm by Ondov *et al.* to identify significant containments of single

genomes within metagenomes [350]. In short, we load all L_i bloom-filters into memory and store a hash-set to represent Q_T. As we stream k-mers from reads or sequences in a given query genome in the same approach as done sub-module 1.1, we add k-mers to the hash-set only if a k-mer is contained within any L_i. After processing all reads or sequences in the query genome, we can use algorithm 1 and its corresponding statistical tests to calculate the fractional genomic containments of each node in the tree, summarized as a tab-separated output.

Redwood2 is implemented in the Scala programming language making use of open-source libraries for the bloom-filter and HyperLogLog algorithms.

5.2.3 Benchmarking *Redwood2*

A *Saccharomyces sensu strictu* genus tree was constructed from 195 previously published genomes collected from multiple studies (see Table 6.1). We used *de novo* genome assemblies when available, and alternatively, WGS-Illumina datasets of at least 10x coverage after adapter and low-quality bases from raw-reads were trimmed with *Trimmomatic* [284]. Bottom-sketches were constructed for each genome using k-mer size of 21, sketch-size of 100,000, and false-positive rate of 1%; for WGS datasets, a minimum k-mer count of 3 or 5% of total coverage (which ever was higher) was used to remove erroneous k-mers (e.g. for a sequencing dataset with 100x coverage, the minimum k-mer count would be set to 5). The HyperLogLog algorithm for approximate node-cardinality was used using a standard error-rate of 1% (corresponding to 14-bit precision value for the counting registers).

Simulated hybrids were constructed by concatenating reference genomes from difference *Saccharomyces* species. In each simulation, one genome was reduced to different-random-relative-proportions while keeping the remaining genomes as is. The reductions was done by an in-house script using the Scala programming language which takes all contigs in an assembly, randomly permutes their order, and outputs contig-sequences until the total number of bases is $r \cdot g$, where r is the desired relative fraction of the original genome and g is the original genome-size. The *expected* fractional genomic contribution for the species of one simulated hybrid is thus calculated as:

$$E[FGC(N)] = \begin{cases} \frac{r}{(n-1)+r}, & \text{if } N = \text{ species reduced} \\ \frac{1}{(n-1)+r}, & \text{o.w.} \end{cases}$$

Where N is a species-node in the *Saccharomyces sensu strictu* genus tree and n is the total number of genomes in the simulated hybrids.

Eight previously published *Saccharomyces* hybrid-genomes were used to additionally validate *Redwood2* (see Table 5.2). *De novo* assemblies were used for *S. cerevisiae x S. eubayanus* and *S. cerevisiae x S. kudriavzevii* (as they were available) and Illumina datasets were used in the remaining hybrid-genomes after processing raw-reads as previously described.

Table 5.1: Genomes used to construct the *Saccharomyces sensu strictu* tree. The genomes are a mixture of read-sets and *de novo* assemblies based on the available sequencing information. For the same reasons, only a limited number of genomes for each species were used.

Species	Strains		Reference
S. arboricola			
	H-6		[338]
	Total	1	
S. cerevisiae			
	Total	165	[15]
S. eubayanus			[265]
	yHRVM107		[265]
	yHKS212		[265]
	yHCT104		[265]
	yHCT63		[265]
f	yHDPN422		[265]
	yHKS509		[265]
	yHKS211		[265]
	yHCT101		[265]
	yHCT61		[265]
	yHCT72		[265]
	yHDPN424		[265]
	yHKS210		[265]
	yHDPN421		[265]
	yHCT99		[265]
	yHRVM108		[265]
	CBS12357		[265]
	yHDPN423		[265]
	yHCT96		[265]
	Total	18	[265]
S. jurei			
	SacJureiUoM1		[336]
	Total	1	
S.kudriavzevii			
	CR85		[337]
	SKCA111		[337]
	Total	2	
S. mikatae			
	IFO1815		[354]
	Total	1	
S. paradoxus			
	CBS432		[260]
	N44		[260]
	UWOPS919171		[260]
	UFRJ50816		[260]
	Total	5	
S. uvarum			
	CBS7001		[355]
	yHCT77		[264]
	CRUB1994		[264]
	CRUB1776		[264]
	CRUB1990		[264]
	CRUB1989		[264]
	CRUB1987		[264]
	DBVPG7787		[264]
	CRUB1988		[264]
	CRUB1984		[264]
	GM14		[264]
	CRUB1993		[264]
	CRUB1991		[264]
	yHCT78		[264]
	Total	14	

Table 5.2: Published *Saccharomyces* hybrid-genomes used to benchmark *Redwood2*. The set of eight strains are a diverse collection of *Saccharomyces* hybrids described in their respective studies as either beer, wine, or cider fermenting isolates. These genomes were used to benchmark *Redwood2*'s capability in inferring species composition in real-world hybrids.

Hybrid	Strain	Reference	Data-type
S. cerevisiae x S. eubayanus			
	WE34_70	[208]	*De novo* assembly
	CBS1538	[208]	*De novo* assembly
	CBS1483	[212]	*De novo* assembly
S. cerevisiae x S. eubayanus x S. uvarum			
	NCAIM676	[264]	Illumina reads
S. cerevisiae x S. kudriavzevii			
	VIN7	[340]	*De novo* assembly
	HA1836	[356]	*De novo* assembly
S. cerevisiae x S. kudriavzevii x S. uvarum			
	CID1	[264]	Illumina reads
	CBS2834	[264]	Illumina reads
S. cerevisiae x S. uvarum			
	Muri	[263]	Illumina reads

5.3 Results and discussion

We developed *Redwood2* to estimate the global *Saccharomyces* species composition in a *de novo* assembly or whole-genome sequencing dataset, facilitating downstream evolutionary analysis of hybrid *Saccharomyces* genomes. This computational problem is inherently the same as inferring the presence and abundance of taxa (e.g. species, genus, and family content) in a metagenomic dataset. One popular approach is to derive k-mers from large large public databases of microbial genomes in conjunction with inferred phylogenetic relationships to enable taxonomic classifications [357]. Our proposed method also follows a k-mer based approach, and adapts probabilistic genome representations to make use of growing public datasets of *Saccharomyces* genomes. We first construct a (phylo-)genetic tree of 195 *Saccharomyces* genomes covering all major species in the *Saccharomyces sensu strictu*. Using this tree, we define that the genomic contribution of a node (e.g. species, lineage, strain) from a tree in a query genome can be seen as, $FGC(N) = \frac{\Theta_{U_N}}{\Theta_{R_N}}$, where Θ_{U_N} is the number of unique k-mers uniquely contained in node N or any of its children, while Θ_{R_N} is the total possible number of unique k-mers from the query genome that could be assigned to node N. This exact calculation requires one to test the membership of all k-mers in a whole-genome sequencing dataset to all nodes in the tree—a calculation that is computational feasible when using probabilistic genome representations. As such, we adapt recent work on probabilistic data structures (e.g. genome sketches, sequence bloom-trees, and the HyperLogLog algorithm) to enable fast calculation of $FGC(N)$. We benchmarked *Redwood2* on simulated and real *Saccharomyces* genomes, showcasing its ability to quickly and accurately determine species composition in a whole-genome sequencing dataset of a *Saccharomyces* isolate.

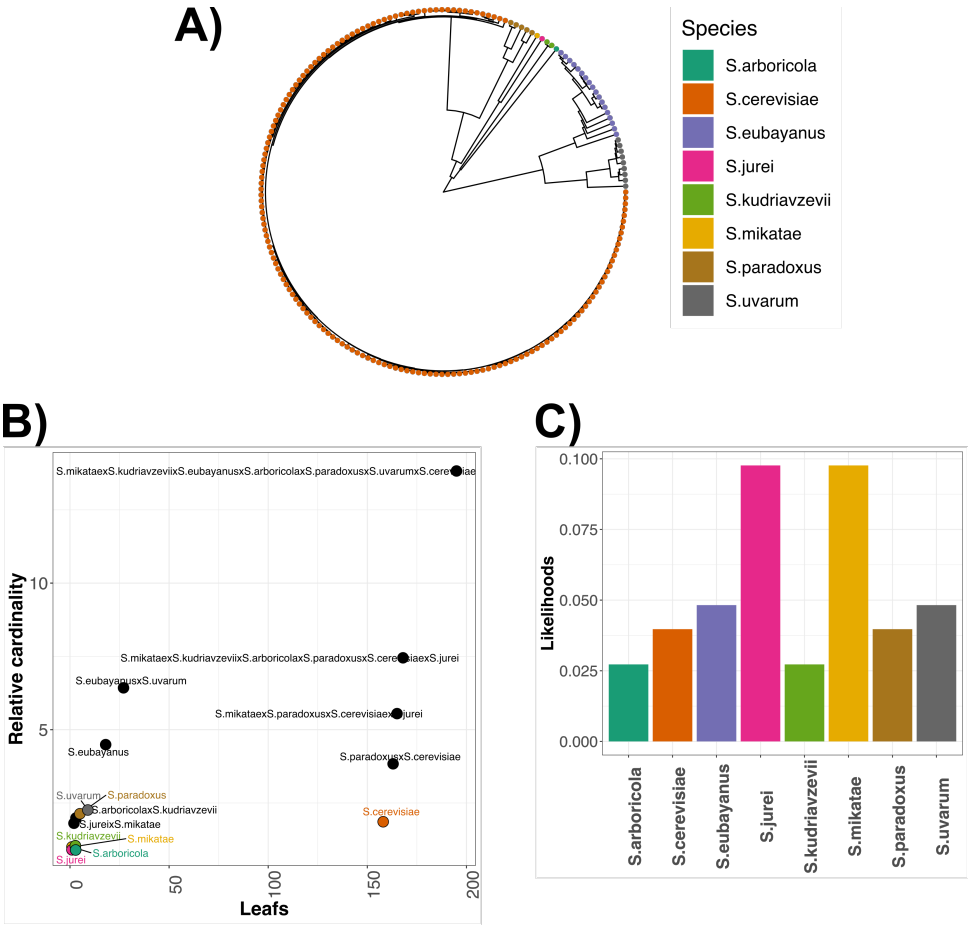


Figure 5.2: ***Saccharomyces sensu strictu* tree as constructed by *Redwood2*.** (A) Hierarchical clustering of 195 genomes from eight different *Saccharomyces* species. (B) Node cardinalities (total number of unique k-mers) of different parent nodes in the tree relative to initial sketch-sizes of 100,000. (C) The likelihood that a k-mer from some parent-species node from the tree shown in (A) is also present in any other parent-species node based on the relationships defined in the tree shown in (A). All colors (with the exception of black) correspond the 'Species' color-legend to the right of (A).

5.3.1 *Saccharomyces sensu strictu* tree construction

We first constructed a partial sequence bloom-tree of the *Saccharomyces sensu strictu* group. Ideally, one would construct a tree using hundreds or thousands of samples uniformly spanning the eight *Saccharomyces* species, enabling *Redwood2* to better estimate the genomic contribution due to a large reference dataset. However, there is an in-balance in the number of publicly available genomes for each species. For example, *S. cerevisiae* has over 1,000 characterised public genomes [15, 73], contrast to *S. jurei*, *S. kudriavzevii*, *S. arboricolus*, *S. mikatae*, and *S. paradoxus* where only 1-13 genome are publicly available [260, 336–338, 354]. At the time when this study was conducted, *S. eubayanus* and *S.*

uvarum each had less than ~60 public genomes available [264, 265]. Nevertheless, we used 195 total *Saccharomyces* genomes to represent *Saccharomyces sensu strictu* group, consequently reflecting the in-balance of data availability for each species (see Figure 5.2 A).

Figure 5.2 A shows a hierarchical clustering of the 195 *Saccharomyces* genomes, and accurately reflects the evolutionary histories of *Saccharomyces sensu strictu* species [336, 358–360]. Additionally, the genomic diversity observed in each of the species varies. Figure 5.2 B shows the relative cardinalities (that is, the total number of unique k-mers) for the parent-node of each species using sketch-size of 100,000. Disregarding hybrids, *S. cerevisiae* is the least genomically diverse population with a cardinality 1.86x of the original sketch-size, followed by *S. eubayanus*, *S. uvarum*, and *S.paradoxus*. We cannot infer the genetic diversity for species-nodes with only one sample, such as *S. jurei*, *S. kudriavzevii*, and *S. mikatae*. These results imply different sensitivity levels when estimating genomic contribution for each species since *Redwood2* infers statistical significance in the estimations based on the likelihood of a k-mer being shared between two different species based on the defined linear relationships in a phylogenetic tree.

Figure 5.2 C shows the likelihood that a random k-mer contained in one species-node will be shared with any other species-node in the tree. For example, given a random set of k-mers drawn from the *S. cerevisiae* species node, 4.82% of those k-mers are expected also be present in *S. eubayanus*. The likelihoods thus indicate the minimum level of detection for each species. *S. arboricola* and *S. kudriavzevii* harbours the lowest likelihoods at 2.7%, while both *S. jurei* and *S. mikatae* have the highest at 9.77%. The former results are as expected given the relationships defined in the tree in Figure 5.2 A, where the genomes for *S. arboricola* and *S. kudriavzevii* are largely distinct single, branching out-groups relative to the other species, implying a higher fraction of unique k-mers in their genomes, and hence a lower fraction of shared k-mers with the other species (see Figure 5.2 A and C). *S. jurei* and *S. mikatae* are relatively more similar in their k-mer content and appear to have a higher fraction of shared k-mers with their closest neighbour, the parent node of *S. cerevisiae* and its closest wild relative, *S. paradoxus*, which has been similarly reported by [336].

5.3.2 Redwood2's estimated species contributions are accurate in a simulated benchmark

To evaluate *Redwood2*'s accuracy in estimating global species contribution in *Saccharomyces* hybrid genomes, we simulated double and triple-hybrid genomes with various level of species contribution (see Figure 5.3 A). More specifically, we varied the relative fraction of genomic contribution (FGC)—that is, the fraction of genomic content originating from some genome—for each simulated hybrid, reflecting known reported cases of natural *Saccharomyces* hybrid genomes involved in beer, wine, and cider fermentation (see Table 5.2). For example, group-2 *S. pastorianus* genomes (a hybrid of *S. cerevisiae* and *S. eubayanus*) can harbour the majority of chromosomes from both species in its nucleus [208, 212], which would lead to an FGC value of 50% for each species. Contrast to group-1 genomes where only a small fraction of *S. eubayanus* chromosomes remain [208, 212], which would lead to proportionally higher FGC values for *S. cerevisiae* and lower values for *S. eubayanus*. For the triple-hybrids, at least one species was fixed to retain all genomic content in the simulated triple-hybrids, which would lead to a maximum FGC value of 50%.

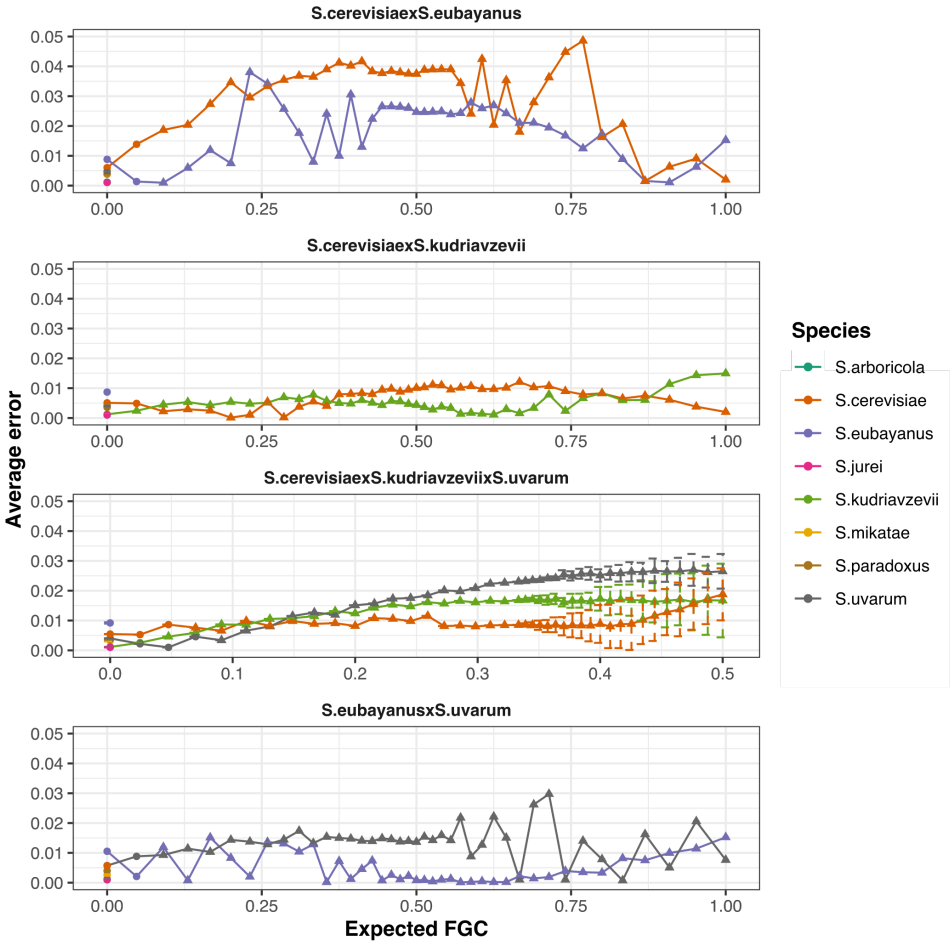


Figure 5.3: Benchmarking *Redwood2* using simulated hybrid-genomes and a real global population of *S. cerevisiae* strains. (A) The average error (y-axis) in *Redwood2*'s estimated species contribution (referred to as FGC; see methods for exact calculation) for four different simulated hybrid genomes as a function of the expected species contribution (x-axis). For each estimation, triangles represent statistical significance (p-val ≤ 0.05) after Bonferroni correction. These results indicate high accuracy of ≥ 95% in a simulated setting involving *Saccharomyces* hybirds involved in beer, wine, and cider fermentation. (B) Correlation of *Redwood2*'s estimations of the species contribution from *S. paradoxus* and *S. cerevisiae* across 1,011 *S. cerevisiae* global strains from [73]. Despite the global genomic diversity in the collection of *S. cerevisiae* genomes, *Redwood2* is still able correctly estimate a *S. cerevisiae* species contribution of ≥ 99% for the vast majority of strains (bottom-right in figure). In cases where the estimation is < 99%, we observe an increase in the estimated species contribution from *S. paradoxus*, suggesting partial *S. paradoxus* alleles in the genomic dataset. All estimations were based on the tree constructed in Figure 5.2A.

We observed a 0-5% error in *Redwood2's* estimated global species composition (see Figure 5.3 A). For example, in the simulated *S. cerevisiae x S. eubayanus* genomes, the error for both species ranges from 0.09% to 4.86% (see Figure 5.3 A). Only the estimation at the expected FGC value of ≤ 4.76% for both species was deemed in-significant (i.e. p-values ≤0.05 after multiple testing correction), reflecting the minimum level of detection as discussed in section 5.3.1 (see Figure 5.2 A). Additionally, the calculated errors for species not in the simulated genome (e.g. false-positive species calculations) is minimal, with a maximum of 0.47%, and all p-values are > 0.05 after multiple-testing correction, indicating high specificity to discern *Saccharomyces* species not present in a given genome (see Figure 5.2). In the *S. cerevisiae x S. kudriavzevii* hybrid simulation, we observe lower errors ranging from 0.01% to 1.21%, and all expected values > 0% are deemed significant except at 4.76% in *S. cerevisiae*, with no significant contributions from other species (see Figure 5.2 A). These results show that *Redwood2* can accurately approximate the global species composition in (simulated) hybrid *Saccharomyces* genome.

The results shown in Figure 5.3 A additionally suggests that a relatively higher fraction of k-mers are shared by *S. cerevisiae* and *S. eubayanus*, despite previous findings that *S. cerevisiae* and *S. eubayanus* are phylogenetically more distant than *S. cerevisiae* and *S. kudriavzevii* [338, 361, 362]. For example, *Redwood2* has a relatively high error in estimating the species composition in a *S. cerevisiae x S. eubayanus* hybrid than a *S. cerevisiae x S. kudriavzevii*. We hypothesize that this discrepancy may be due differences in the sequencing datasets used for both the construction of the *Saccharomyces sensu strictu* tree and the simulated hybrid genomes. For example, the *S. cerevisiae* and *S. kudriavzevii* genomes were based on short-read Illumina assemblies (see section 5.3.2), and hence the k-mer content originates from consensus representations of heterozygous regions (if diploid) and collapsed repeats [15]. The majority of *S. eubayanus* and *S. uvarum* datasets used in the tree construction exists only as read-sets, and hence the k-mer contents also include heterozygous sequences and true repeat-content [264, 265]. As we used a complete *S. cerevisiae* reference assembly to simulate the hybrids, k-mer content from typically collapsed regions (e.g. sub-telomeric and ribosomal loci [207, 212, 260]) may yield false-positive hits to *S. eubayanus* species, explaining the relatively high errors in the *S. cerevisiae x S. eubayanus* hybrid. Nevertheless, we still observed only a maximum of ~5% error in the simulations of these genomes (see 5.3).

A similar performance can be seen in the simulated *S. cerevisiae x S. kudriavzevii* hybrid and the *S. cerevisiae x S. kudriavzevii x S. uvarum* triple-hybrid, with no significant contributions from any other species (see Figure 5.3 A). In particular, the standard deviations for the simulated triple-hybrid increase at an expected FGC value of ~30%, indicating different sensitivity values depending on the mixture of genomic content from different species.

Overall, these results indicate that *Redwood2*'s estimations of the global species composition for a genome is ≥ 95% accurate in a simulated setting involving three *Saccharomyces* species. However, the simulated hybrids were constructed with genomes already included in the *Saccharomyces sensu strictu* tree, and hence, the results described above are biased as they do not assess *Redwood2*'s accuracy when facing genomic diversity absent in the reference tree.

We therefore investigated how genomic diversity absent in a given partial sequence

bloom-tree influences *Redwood2*'s fractional genomic contribution calculation. We opted to use a published dataset of 1,011 *S. cerevisiae* genomes capturing the global diversity of the *S. cerevisiae* species [73], as opposed to simulating genomic diversity *in-silico*. We found that the vast majority of the FGC values for the *S. cerevisiae* species were > 99%, and no sample had statistically significant contributions from any other species (see Figure 5.3 B). For a subset of strains where the estimation is < 99%, we observe a correlation (r = -0.88) between decreasing FGC values of *S. cerevisiae* and increasing values for *S. paradoxus*, suggesting partial genomic sequences from the *S. paradoxus* species, which is the closest wild relative to *S. cerevisiae* [225, 260]. Interestingly, the six outliers in the bottom left corner of Figure 5.3 B (i.e. $FGC(S.cerevisiae) < 95.4\%$ and $FGC(S.paradoxus) < 0.03\%$) are the strains *AMH*, *BAL*, *CDH*, *CEG*, *CEI*, and *CFH*—which are among the top most divergent strains based on the total number of single nucleotide variants as reported in their respective study *et. al* [73]. Although there are small predicted genomic contributions from *S. paradoxus* as well as *S. eubayanus* in these strains, these contributions are minimal and deemed insignificant (p-value > 0.05), suggesting that the genomic sequences of these strains are not entirely of *Saccharomyces* origin. Indeed, *CDH* and *CFH* were reported to have genomic contamination from *Staphylococcus epidermis* [73], explaining the relatively lower species contribution from *S. cerevisiae* by *Redwood2*, and it may be possible that the other strains similarly harbour non-*Saccharomyces* originating sequences. Nevertheless, these results show minimal influence in *Redwood2*'s species estimation when faced with unobserved genomic diversity, at least when the diversity is in similar range to that of the global *S. cerevisiae* genomic diversity.

5.3.3 *Redwood2* provides informative global species estimations in public hybrid genomes

In addition to the simulated benchmark described in the previous section, we also evaluated *Redwood2*'s accuracy in estimating the global species composition in real published hybrid-genomes (see Figure 5.4 and Table 5.2). In general, *Redwood2's* estimations across the different hybrid-genomes reflect the evolutionary histories reported in their respective published studies (see Figure 5.4), and showcase the applicability of *Redwood2* for studying the evolutionary histories of different *Saccharomyces* hybrids. For example, CBS1483 and WE34_70 are group-1 *S. pastorianus* strains (a hybrid of *S. cerevisiae* and *S. eubayanus* with slightly higher *S. cerevisiae* genomic content due to higher chromosome copy-numbers [208, 212]. This is contrast to the Group-1 originating strain, CBS1538, which retains only a few *S. eubayanus* chromosomes. Indeed, *Redwood2's* estimations precisely reflect the genomic contributions from the two evolutionary groups (see Figure 5.4).

Importantly, the ability to systematically estimate the species contribution in *S. pastorianus* genomes can help disentangle two competing theories regarding the evolutionary origins of this species: all *S. pastorianus* genomes originate from a single evolutionary event and a population bottleneck led to two evolutionary groups [208, 212], or there were two separate hybridization events the led to two groups [208]. A hallmark feature distinguishing both groups is the relative genomic contribution from the *S. cerevisiae* and *S. eubayanus* species [208, 212]. Unfortunately, most comparative genomic studies of *S. pastorianus* have only involved a few number of isolates, so the global spectrum of species composition in this species is unknown [208, 212]. As such, a systematic evaluation of the

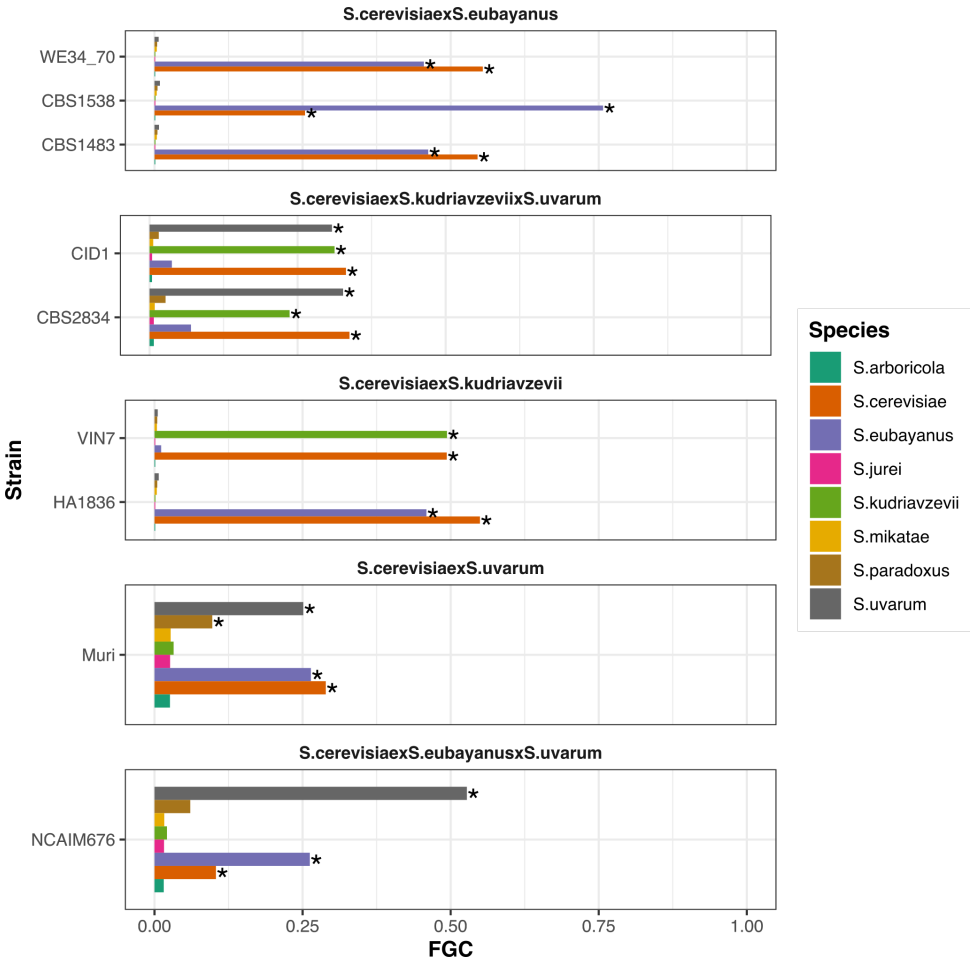


Figure 5.4: *Redwood2* species composition estimations on published *Saccharomyces* hybrid-genomes. The genomes are a mixture of real-world *Saccharomyces* hybrid genomes used in beer, wine, and cider fermentation, whose evolutionary history and species composition has been investigated in their respective studies (see Table 5.2). Each sub-plot represents a distinct *Saccharomyces* species-hybrid with the strain identifiers on their respective y-axis, and bar-plots representing *Redwood2*'s estimated species composition color-coded to represent the eight major species in the *Saccharomyces sensu strictu* group (see Figure 5.1). Asterisks indicate that the corresponding *FGC* calculation is statistically significant (p-val ≤ 0.05) after multiple-testing correction.

species contribution in hundreds to thousands of global *S. pastorianus* isolates can help provide evidence for one of the two competing theories by indicating either a binary or gradient separation in the relative species composition of *S. cerevisiae* and *S. eubayanus* in *S. pastorianus* isolates.

Two additional genomes that we evaluated with *Redwood2* was the *S. cerevisiae x S. kudriavzevii x S. uvarum* triple-hybrid strains, *CID1* (a cider yeast isolated in France) and *CBS2834* (a fendant-wine yeast isolated in Switzerland) [264]. The species composition

for both strains were determined by mapping reads to a concatenated reference of coding sequences from the reference assemblies of *S. arboricola*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, and *S. uvarum*, and resolving multi-mappings reads using divergence thresholds [264]. As such, the reported genomic composition for *CID1* is 36.5% fo S. cerevisiae, 33.5% for S. kudriavzevii, and 29.6% for S. uvarum. The reported genomic compositions for *CBS2834* are 36.2%, 28.1%, and 30.2%, respectively. Indeed, *Redwood2*'s *Saccharomyces* species estimations for both genomes precisely reflect the reported compositions using the same sequencing read-set: 33.2%, 31.2%, and 30.8% for *CID1*, respectively; for *CBS2834*, they are 33.8%, 23.6%, and 32.7%, respectively. These results not only showcase *Redwood2*'s accuracy, but also it's feasibility in determining species composition in *Saccharomyces* hybrids, as they do not require choosing a single "best" reference genome with additional evolutionary parameters, especially when one considers the standing genomic diversity within various *Saccharomyces* species [15, 73, 264].

We similarly evaluated two *S. cerevisiae x S. kudriavzevii* hybrid-strains, *VIN7* (a wine-yeast from South Africa) and *HA1836* (a wine yeast from an Austrian vineyard) [340, 356]. *VIN7* was reported to contain a 2:1 ratio in the *S. cerevisiae* and *S. kudriavzevii* genomic contributions, respectively [340]. However, this was based on estimated chromosome copy-number ratios as it was observed to have a homozygous *S. kudriavzevii* sub-genome [340]. Given that we were only able to process *de novo* assembly of this strain (raw-read data was not publicly available) which is a consensus representation of all true haploid chromosomes, *Redwood2* estimated 50% contribution from each of the two species (see Figure 5.4). Interestingly, although *HA1836* was reported to have a similar composition to strain *VIN7* [356], *Redwood2* indicates that *HA1836* is a *S. cerevisiae x S. eubayanus* hybrid-strain with a slightly higher *S. cerevisiae* sub-genome (see Figure 5.4). Based on the relatively higher genome content of *VIN7* along with our results regarding *S. cerevisiae* and *S. eubayanus* hybrids, *VIN7* would be classified as a group-2 strain. Indeed, the estimated sequence divergence based on the native implementation of the MASH-algorithm [177] of *HA1836* to the group-2 strains, *CBS1483* and *WE34_70*, is 7.44e-04 and 9.82e-04, respectively; contrast to the estimated sequence divergence of 0.037 when compared to *VIN7*. As such, *HA1836* is likely a *S. cerevisiae x S. eubayanus* hybrid as opposed to the reported *S. cerevisiae x S. kudriavzevii* hybrid, further showcasing the applicability of *Redwood2* to objectively study hybrid *Saccharomyces* isolates.

The *S. cerevisiae x S. uvarum* hybrid-strain, *Muri* (a beer yeast from Norway), was reportedly characterized to mostly harbour *S. cerevisiae* and *S. uvarum* chromosomes, with significant chromosomal introgressions from *S. eubayanus* [263]. Although the exact estimated compositions were not calculated, Krogerus *et al.* visually report a full *S. cerevisiae* sub-genome and a hybridized version of *S. uvarum* and *S. eubayanus* chromosomes [263]. *Redwood2* reports genomic contributions of 28.8%, 25.1%, and 26.4% for *S. cerevisiae*, *S. uvarum*, and *S. eubayanus*, respectively. Additionally, *Redwood2* reports 9.77% genomic contribution from *S. paradoxus*. To verify the accuracy of the *S. paradoxus* estimate, we aligned reads to a concatenated reference of *S. cerevisiae S. eubayanus*, and *S. uvarum* as described by [263], and additionally added the *S. paradoxus* genome, CBS432 [260]. Using very conserved thresholds to curate the multi-mapping reads, we found ~123,000 unique read-alignments to *S. paradoxus*, but the vast majority of these alignments were mapping to the sub-telomeric regions. In a similar manner to the species estimation errors induced

by using full short-read Illumina datasets (see section 5.3.2), *Redwood2*'s species estimation of 9.77% for *S. paradoxus* is likely a false-positive estimation arising from genomic sequences in the sub-telomeric regions that are only accessible in the long-read PacBio assemblies, contrast to the rest of the short-read Illumina assemblies used to construct the *Saccharomyces sensu strictu* tree [207, 212, 260]. Despite the fact that some samples in the tree are based on Illumina read-sets containing sub-telomeric regions, these regions are often hyper-variable and differ across *Saccharomyces* genomes [207, 214, 260]. Thus, more uniform representation of sub-telomeric sequences across the tree would be required to properly assess k-mers deriving from these regions.

Finally, the *S. cerevisiae x S. eubayanus x S. uvarum* triple-hybrid (also known as *S. bayanus*) strain, *NCAIM676* (isolated from an unknown fermented drink in Hungary), was reported to have a species composition with 72.2% *S. uvarum*, 26.8% *S. eubayanus*, and 0.98% *S. cerevisiae* [264]. Using the same read-set, *Redwood2* estimates follows the same pattern but with different contributions: 52.7% *S. uvarum*, 26.2% *S. eubayanus*, and 10.4% *S. cerevisiae*—the remaining 10% comprising of statistically insignificant contributions from the other species (see Figure 5.4). The reported species composition of *NCAIM676* was determined using the same method as *CID1* and *CBS2834*, but with a second-round of alignment-filtering to distinguish sequences deriving from *S. eubayanus* or *S. uvarum* due to their high-sequence homology [264]. Although *Redwood2's* estimation are relatively proportional, the reported differences may be due to inherent biases in their methodologies.

Overall, *Redwood2* processed *de novo* assemblies in less than 55 seconds with no more than 512 Mb of RAM using a single thread, and 1.7 Gbp Illumina read-set (e.g. 141x coverage for 12 Mbp genome) in 5:38 minutes with four threads using no more than 769 Mb of RAM—both on a standard Mac laptop (i.e. 2.3 GHz Dual-Core Intel Core i5).

5.3.4 Redwood2 limitations

Redwood2's accuracy is ultimately influenced by the quality of the constructed partial sequence bloom-tree. As discussed in section 5.3.1, there is an in-balance in the number of publicly available genomes for the members of the *Saccharomcyes* species, and hence a biased sampling of the genomic diversity used by the current our constructed *Saccharomyces sensu stricto* tree. A future build could integrate an additional ~200 WGS dataset of *S. eubayanus* from a recent study [363]. Nevertheless, the current tree provides accurate approximations even for species where only a limited number of samples are available. Importantly, the accuracy in the species assignment for the genomes used construct the partial sequence bloom-tree is critical, as mislabelling could lead to incorrect species-to-k-mer assignment. *Redwood2* is also influenced by the sketch-sizes used to construct the partial bloom-tree. In this study, we used sketch-size of 100,000, as we found it to be a size that minimises computational resources but retains relevant sensitivity levels to identify statistically significant species contributions in (natural) hybrid-genomes (see Figure 5.4). Additionally, false-positive allocation of unique k-mers to nodes in the tree should increase when a large number of k-mers from a query genome are streamed (e.g. high-coverage read-set relative to a *de novo* assembly). This is observed in Figure 5.4 A, where queried read-sets have proportionally higher false-positive allocations to unexpected species relative to queried *de novo* assemblies. Decreasing the false-positive rate during tree con-

struction can reduce false-positive hits in high-coverage read-sets, at the cost of increase memory requirements.

Discrepancies between reported and estimated global species composition in real *Saccharomyces* genomes additionally highlight differences in accessibility of genomic information in the current *Saccharomyces sensu stricto* tree. As shown in Figure 5.4 B, a small fraction of the genomic contribution is incorrectly assigned to *S. paradoxus*, albeit not statistically significant. The *S. cerevisiae* genomes used to construct the *Saccharomyces sensu strictu* tree are based on *de novo* assemblies from Illumina reads [15], contrast to the *S. paradoxus* genomes which are *de novo* assemblies from PacBio data [260]. Given that long-read assemblies better assemble repetitive regions (e.g. sub-telomeric genes) [207, 212, 260], the sketches of the *S. paradoxus* genomes could contain k-mers derived from homologous sub-regions that are present in the *S. cerevisiae* assemblies, but not fully represented due to missing/collapsed repetitive regions. A similar situation could occur when comparing *de novo* assemblies and short-read datasets, as assemblies can fail to capture both repetitive sequences and heterozygous information since they are largely consensus representations of the true genome. Indeed, we observe higher errors in the simulated *S. cerevisiae x S. eubayanus* hybrid which involves a high-quality *S. cerevisiae* assembly with closed gaps and Illumina read-sets of the *S. eubayanus* population (see section *benchmark*). Similarly, we also observe relatively higher *S. eubayanus* species estimation in the two *S. cerevisiae x S. kudriavzevii x S. uvarum* triple-hybrids (see Figure 5.4). Although these issues may not lead to statistically significant species estimations, they affect the overall proportions of the estimated species composition values (e.g. summing less than 100%).

5.4 Conclusion

Redwood2 is a fast and alignment-free approach to estimate the global species composition of one or more species in a *de novo* assembly or whole-genome sequencing dataset of a *Saccharomyces* isolate. Our method can thus facilitate evolutionary investigations of (large) collections of *Saccharomyces* genomes. In particular, we were able to provide rapid and accurate estimations in the species composition of real-world industrial hybrids involving cider, beer, and wine fermentation. In the latter case, *Redwood2* provided evidence for a mis-classification in the species composition of an existing wine strain.

6

Approximate, simultaneous comparison of microbial genome architectures via syntenic anchoring of quiver representations

A long standing limitation in comparative genomic studies is the dependency on a reference genome, which hinders the spectrum of genetic diversity that can be identified across a population of organisms. This is especially true in the microbial world where genome architectures can significantly vary. There is therefore a need for computational methods that can simultaneously analyze the architectures of multiple genomes without introducing bias from a reference.

In this paper, we present Ptolemy: a novel method for studying the diversity of genome architectures—such as structural variation and pan-genomes—across a collection of microbial assemblies without the need of a reference. Loosely speaking, Ptolemy is a "top-down" approach in comparing whole genome assemblies: genomes are represented as labelled-multi-directed graphs—known as quivers—which are then merged into single, canonical quiver by identifying "gene anchors" via synteny analysis. The canonical quiver therefore represents an approximate, structural alignment of all genomes in a given collection encoding structural variation across (sub-)populations within the collection. We highlight various applications of Ptolemy by analyzing structural variation and the pan-genomes of different datasets composing of Mycobacterium, Saccharomyces, Escherichia, and Shigella species. Our results show that Ptolemy is flexible and can handle both conserved and highly dynamic genome architectures. Ptolemy is user-friendly—requires only FASTA-formatted assembly along with a corresponding GFF-formatted file—and resource-friendly—can align 24 genomes in ~10 mins with 4 CPUs and < 2GB of RAM

This chapter has been published in *Bioinformatics* [364]

6.1 Introduction

Single-molecule sequencing technology has enabled near-complete reconstruction of microbial genomes in both bacterial and eukaryotic organisms [146, 207, 259, 260]. Furthermore, ultra-long reads—such as those obtained from Oxford Nanopore Sequencing Technology—can greatly facilitate completion of genome assemblies [147]. This information enables a more comprehensive understanding of the genomic architecture, variation, and evolution of microbial species [207, 259, 260]. As single molecule sequencing technologies become more accessible, high-quality microbial assemblies are expected to become more prevalent, decreasing the dependency of a reference genome in comparative studies and instead shifting towards direct assembly-to-assembly analysis.

In general, comparative genomic studies aim to identify differences and similarities in the genetic content of a collection of genomes. Depending on the nature of the research question, this can be achieved via two strategies: "bottom-up" and "top-down". Bottom-up approaches are essentially (multiple) whole genome alignment which use short sub-sequences to anchor and align genomes and which then undergo (multiple) sequence alignment [241, 365–368]. One classic tool is *MUMmer* [241], which aligns a query genome to a reference genome using maximal unique matches (MUMs). Clustering of MUMs can then highlight structural differences—such as translocation, inversions, large insertions and deletions—between the query and reference [241]. Sequencing projects dealing with collections of (novel) assemblies often use *MUMmer* to align the genomes to a common reference and identify variations across the collection of genomes by globally comparing differences between each query and reference [147, 207, 260, 369]. However, comparatives results can be biased as these variants only account for differences in sequence that is shared between the query and reference genome. More specifically, nested variation—such as unique sequences in a collection of genomes that are absent in the reference but themselves contain additional variation among each other—are missed.

Multiple-whole genome alignment approaches offer higher resolution of nested variation that can exists across a collection of genomes. Tools like the *EPO* pipeline [367], *Cactus* [366], *ProgressiveMauve* [368], and *Mugsy* [365], utilize anchor-sequence-finding methods (e.g. MUMs) across a set of genomes to identify collinear regions and thereafter induce multiple sequence alignments across those regions. These approaches are particularly useful in identifying single nucleotide variants (SNPs) and insertion and deletions (INDELs) across several assemblies without bias of a reference. In particular, *ProgressiveMauve* and *Mugsy* have been designed in the context of microbial assemblies with *ProgressiveMauve* tolerating structural variation—such as inversion—common in microbial species [365, 368]; enabling both sequence and structural variation discovery across a collection of genomes. Nevertheless, a major limitation of these approaches is scalability as they have run-times that can take several hours/days depending on genome divergence [365, 368].

Alternatively, the "top-down" approach in comparing assemblies uses pre-defined biological features as opposed to raw DNA sequence. One widely-studied approach is synteny analysis: using gene annotations to identify sets of (coding) sequences that are similar/different across a set of genomes [370]. The intuition is that (evolutionary) closely related genomes are not random and instead share a similar genomic structure—such as gene order—due to some common ancestor. The aim is then to identify orthologous sequences

across two or more genomes and find segments that maximally extend the collinearity of the gene order, often referred to as synteny. Tools like *i-ADHore* [371], *Proteny* [372], *SynFind* [373], and *SynChro* [374] aim to identify syntenic regions across a collection of two or more genomes which can then be processed down-stream for further characterization. It is important to note that these methods heavily rely on pre-defined gene annotations and are therefore sensitive to annotation errors. Furthermore, syntenic regions are computationally less expensive to compute since the annotations—equivalent to sequence anchors in methods using the bottom-up approach—are pre-defined. Because the goal of these methods is to compare genomes in terms of gene-order and content, the analysis is generally restricted within one or several syntenic regions [371–374].

The use of graph-based data structures for comparing multiple genomes has recently been highlighted. More specifically, the paradigm of computational pan-genomics aims to combine multiple assemblies into a single, graph-based data structure to reduce reference bias and enable more robust analysis of variation that exists within a (sub-)population [174]. The benefit of this approach has been demonstrated in alignment and variant calling analysis of short-read data sets [375, 376]. In these studies, existing variation were integrated into a common reference genome represented as a graph, which facilitated better placements of short-reads to difficult regions (e.g. highly variable regions), providing a better understanding of the allele composition of those regions within (sub-)populations [375, 376].

Implementations of graph-based data structures in comparative genomics is not new and has been previously used for a wide-range of genome analysis applications. In terms of microbial genome comparison, the utilization of graph-based representations have been used to compare multiple genome assemblies using a combination of the "bot-tom-up" (DNA-sequence-based) and "top-down" (synteny/gene annotation-based) approach. *Sibelia*, for example, concatenates multiple genomes sequentially into a single "virtual" genome which is then decomposed into a DNA sequence-based kmer de Bruijn graph [377]. Sets of nodes that are sequentially-identically "labeled" (e.g. kmer sequence) are merged thus leading to an alignment de Bruijn (A-Bruijn) graph data structure [377]. *DRIMM-synteny* [378]—a predecessor of *Sibelia*—uses a similar approach except that it works at the gene-level: nodes are genes, kmers consist of the alphabet of assigned gene labels, and the A-Bruijn graph is constructed by applying the "gluing" operation on identical labeled kmers. Similarly, *Pandaconda* [379] uses pre-assigned family protein labels across multiple genomes, decomposes the genomes into a de Bruijn graph, and applies the gluing operation on identically labeled nodes. Therefore, genetic variation—encoded as alternate paths of genes and gene families—highlight architectural differences across multiple genomes. A major difference is that *Pandaconda* does not modify the graph to remove cycles enabling discovery of complex structural variations across a set of genomes [377–379]. It is also important to note that both *DRIMM-synteny* and *Pandaconda*—which used the "top-down" approach—require pre-assigned labels such that genes that are considered to be identical (e.g. orthologous) have the same label [378, 379]. Ultimately, these graph-based approaches aim to summarize the genetic content of multiple genomes in a single graph data structure to identify genetic variation across multiples assemblies; attempting to place biological context surrounding variation that exists across the genomes.

Here, we present *Ptolemy*: a method to simultaneously compare the genome archi-

tectures of collections of microbial assemblies using both gene synteny and sequence information. *Ptolemy* is a graph-based and gene annotation approach to aligning multiple genomes (e.g. "top-down"), similar to the A-Bruijn methods previously mentioned. However, *Ptolemy* does not require pre-assigned gene labels and instead computes these labels by identifying maximally-syntenic-ortholog-clusters of sequences based on the corresponding gene annotations of an assembly. Furthermore, *Ptolemy* represents the assemblies via a labelled-multi-digraph model (also known as quivers) and uses subsequent morphism mappings to align multiple genomes into a canonical quiver. The resulting representation thus captures structural across a collection of genomes into a single graph data structure which can then be extracted using dynamic maximally-labelled path traversal and intuitively visualized with available graph visualization software.

6.2 Methods

The algorithms for our graph and synteny-based approach for simultaneous alignment of multiple genomes is packaged into Ptolemy and takes as input a set of FASTA-formatted assemblies along with their gene annotations in GFF-format. The two novel contributions of *Ptolemy* are the genome representation and corresponding utilities of labelled-multi-directed graph (also known as quivers) and the syntenic-anchor finding algorithm. In the following sections, we provide a detailed description of the algorithms used in *Ptolemy*: first, we describe the quiver representation of a genome and morphism mappings to structurally align multiple genomes without the need of a reference via a "top-down" approach (e.g. orthologous genes). We then describe our implementation of construct-ing such representation using syntenic-anchors based on syteny-based ortholog clustering. Finally, we describe how structural variation can be extracted from the quiver as a population using dynamic path traversal of labelled edges.

6.2.1 Synteny and the quiver representation of genomes

As previously mentioned, synteny analysis exploits the property that the locations of genes in evolutionary close genome are not random but instead share common structures such as gene order [370–374, 380, 381]. The term ortholog has been used to describe gene sequences between two genomes that derived from a common ancestral gene due to strain/species deviation [381, 382]. Intuitively, two closely related genomes will retain a large fraction of orthologs along with the order of which they appear throughout the genome, referred to as synteny [383]. Overtime, structural variation (such as gene duplications and deletions) and chromosomal rearrangements (including translocations, inversions, and horizontal gene transfer) disrupts synteny between genomes [383]. These disruptions are therefore indicative of structural variation [379, 383]

Under the context of a directed graph, the disruption of synteny would induce alternate paths between genomes. Let a genome, G, be represented as a graph, $G = (V, E)$, where the vertex set V contains all genes in it's respective genome. The edge-set, E, is a set of directed edges describing the sequence of adjacent genes as they appear in the genome such that two adjacent genes v and w are connected by a directed edge, e, describing $v \longrightarrow c$. Note that this high-level graph representation of a genome will contain disconnected connected components if multiple chromosomes are present. Now imagine a working

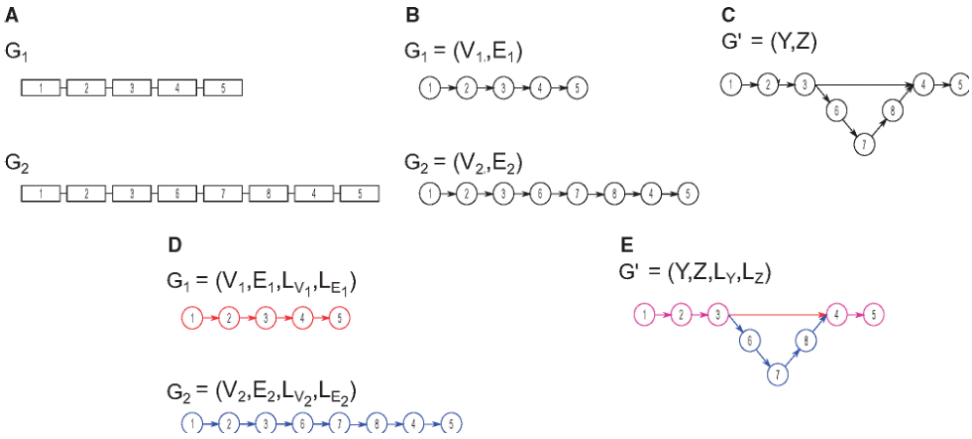


Figure 6.1: Representing genome architectures as graphs. Figure (A) shows two genomes, G_1 and G_2, each containing a single chromosome with 5-8 genes. Figure (B) shows graphical representation of genomes G_1 and G_2. Merging similar nodes in the genome graphs shown previously results in a new graph, Figure (C). The addition of labels to nodes and edges results in a labelled-multi-directed graph, also known as quivers. Figure (D) shows the quiver representation for genomes G_1 and G_2—in this case the colours corresponds to the labels of G_1 and G_2. Merging of the two quiver similarly results in Figure (E), the canonical quiver representation of genomes G_1 and G_2.

example of two closely related genomes, G_1 and G_2 (see Figure 6.1). Constructing a high-level representation of both genomes will yield nearly identical graph with the exception of topological differences associated with structural variation (Figure ??B). By merging identical nodes and edges—which in this context corresponds to orthologous genes in genomes G_1 and G_2—we create a single, canonical genome graph, G', for both genomes, naturally inducing alternate paths reflecting structural variation (see Figure 6.1C).

The addition of labels to nodes and edges to nodes and edges results in a labeled multi-directed graph, known as a quiver (Figure 6.1D). A *quiver* of genome G, is a graph, $G = (V, E, L_V, L_E)$, where V and E are defined as before, L_V is a function mapping a vertex v to a family of set labels, $\Sigma \mid x \in X$, such that $L_V : v \longrightarrow \Sigma_x \mid \forall v \in V$, and L_e is a function that maps an edge e to Σ_x such that $L_e : e \longrightarrow \Sigma_x \mid \forall e \in E$. In our working examples of genomes G_1 and G_2, Σ_x would correspond to unique identifiers for each chromosome in a each genome (e.g. G_1-CHRI, G_2-CHRI, G_1-CHRII, G_2-CHRII, $\cdots$). Note than an edge thus has a head an a tail. In other words, for two adjacent nodes v and w with directed edge e describing $v \longrightarrow w$, the tail of an edges, termed e_t, is v and the head of an edge, termed e_h, is w.

Creating a single canonical quiver from two or more quiver representations can be formally described through morphisms. A *vertex-morphism* for a quiver is a function, $M_V : V \longrightarrow Y$, that maps vertices from some vertex set Y to alternate quiver representation, $G' = (Y, Z, L_Y, L_Z)$. Similarly, an *edge-morphism* is a function, $M_E : E \longrightarrow Z$, that maps edges from some edge set E to an edge set Z belonging to the alternative quiver representation G'. Therefore, the applications of M_V and M_E on G_1 and G_2 result in the transformation to a single, canonical quiver, G' (Figure 6.1E). In this context, the canonical quiver G' is a graphical representation containing the synteny disruptions (e.g. structural

variation) in G_1 and G_2; and the morphisms M_V and M_E describe either "unique" genes or the merging of orthologous sequences. Acquiring G' from some set of quivers thus requires the construction of the mapping functions M_V and M_E from a set of given quivers.

We have now described how we can obtain a single canonical quiver G' from a set of individual quiver-genome representations. G' describes disruptions of synteny within a set of genomes which are indicative of structural variation across multiple genomes and can be obtained via the construction of vertex and edge-morphisms (e.g. mapping functions). In the next section, we describe our implementation of constructing these morphisms from a set of genomes through synteny-based ortholog clustering.

6.2.2 Construction morphisms via syntenic anchors

We can construct the vertex and edge-morphisms for a canonical quiver by performing synteny-based ortholog clustering. *Ortholog clustering* aims to identify sets of corresponding orthologous sequences across a given number of genomes, and is generally obtained through some form of pairwise sequence alignment (either DNA or protein) combined with phylogenetic-inference. For constructing a canonical quiver representation, we require ortholog clusters that are syntenically supported—in other words, sequences that maximize the synteny in the surrounding region of each gene for all genomes in the cluster. We refer to these clusters as *syntenic anchors*. For example, two genes from two genomes may share high sequence similarity and thus form an ortholog cluster. However, the two genes may be located in completely different areas of the genome sharing no synteny in the surrounding regions. In the context of constructing the vertex and edge-morphisms for aligning multiple genomes, we wish to avoid forming these clusters as they will result in spurious connections of dissimilar regions across multiple genomes.

Figure 6.2 gives an overview of our procedure to identifying syntenic anchors. We present a generalized description of our ap-proach, and exact details can be found in Supplementary Methods in [213]. First, we create a database describing the architecture of each genome such as chromosome content including gene sequence and location (see Figure 6.2A). Genes with overlapping open reading frames are merged together into a single "gene unit" whose boundaries are defined by the minimum and maximum coordinates of all overlapping open reading frames. During the database creation, we attempt to identify repeat expansions by identifying connected graphs induced from self-pairwise-gene alignments (see Figure 6.2A) and assign *repeat ranks* describing the order of genes in these regions. We then identify ortholog clusters throughout all genomes in the database by identifying best reciprocal hits (BRHs) through pairwise alignments of the gene sequences for every pair of genomes (see Figure 6.2B).

Syntenic anchors can then be derived from BRHs by scoring the synteny of their neighbouring regions (see Figure 6.2C). Similar to several synteny region finders [371–374], we use a general window scoring approach (such as nearby genes of a given position) as well as independent left and right flanking windows (nearest genes strictly upstream and downstream) which enables us to handle structural rearrangements such as translocations and inversions. We determine whether a BRH is a syntenic anchor by computing a *synteny score* for each window (see Supplemental Methods in [213]). Conceptually, for some defined window size, we iterate through each position upstream and downstream from a BRH and compute the difference between ex-pected and observed synteny based on the

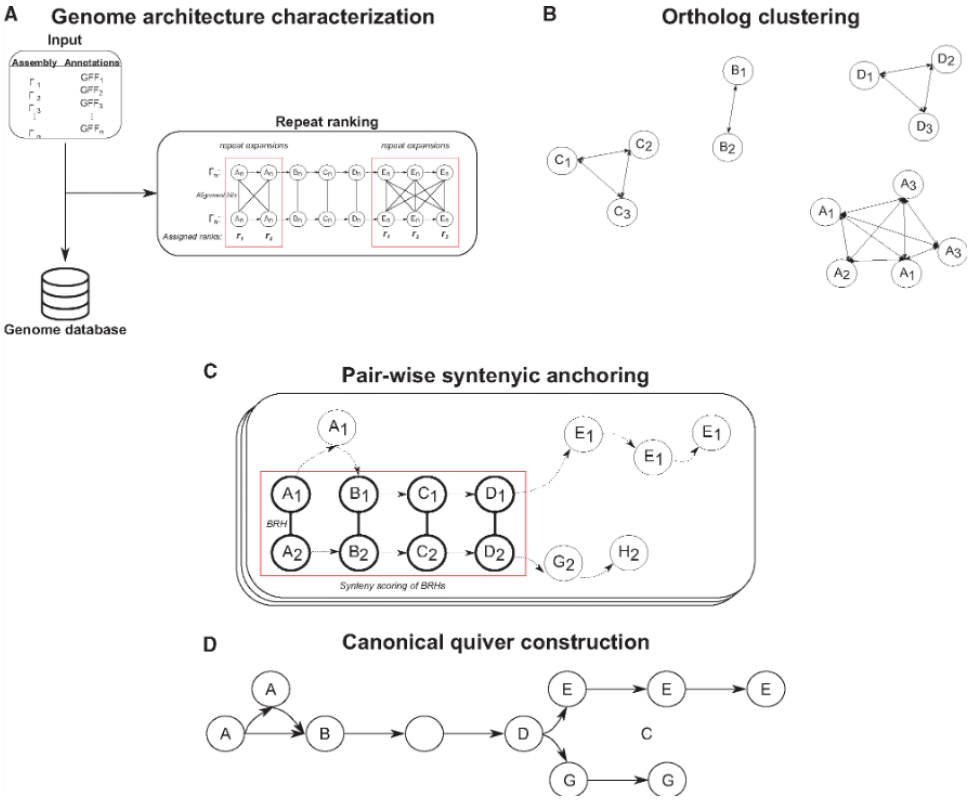


Figure 6.2: Overview of *Ptolemy*. (A) *Ptolemy* first creates a database characterizing individual genome architectures for a given list of assemblies and their corresponding gene annotations. In this process, *Ptolemy* also attempts to identify repeat expansion though self-pairwise gene alignments. (B) Best reciprocal hits (BRHs) are then identified via pairwise gene alignment for every pair of genomes. (C) Syntenic anchors are derive for each BRH by scoring the synteny of the sur-rounding region of corresponding genes. This is done in a pairwise fashion for every pair of genomes. (D) The syntenic anchors are then used to construct the canonical quiver for all genomes in the database.

positional displacement of neighbouring genes (see Supplemental Methods in [213]). In implementation, BRH's are considered syntenic anchors if their synteny score meets a minimum threshold. A detailed description of this parameter along with how to set it can be found in the Supplemental Methods in [213].

Lastly, for a BRH containing genes involved in a repeat expansion, we compute the syntenic score of the neighbouring regions outside of the repetitive region. The intuition is that locally repetitive regions will cause inaccurate calculations for the synteny scores for both genes that are within and around the repetitive region leading to an increase of false negative syntenic anchors. Thus, we "mask" the repetitive regions and compute the synteny upstream and downstream of the region. Furthermore, we restrict the synteny scoring of repetitive genes to those that only have the same *repeat rank* normalizing the syntenic anchors of repetitive regions to their left-most corresponding BRH.

We have described our procedure for identifying syntenic anchors using a synteny-based scoring mechanism for each BRH. The scoring mechanisms accounts for structural variation—such as translocations, inversions, and horizontal gene transfers—and consistently handles repetitive regions such as repeat expansions. With the syntenic anchors in hand, we can construct the edge and vertex-morphisms to create the canonical quiver representation for a given set of genomes. In the next section, we describe our procedure for constructing the morphisms, and hence, the canonical quiver.

6.2.3 Canonical quiver construction

We construct the edge and vertex-morphisms by merging all genes in a syntenic anchor in a single node, implicitly constructing the edge-morphism as well (see Figure 6.2D). Let a sytenic anchor be represented as a family of sets, $A_i \mid i \in I$, where I is the total number of sytenic anchors. By merging all genes in a each A_i, we construct the vertex-morphism, $M_V : v \longrightarrow y \mid \forall v \in A_i, \forall y \in Y$, where Y is the set of nodes in the canonical quiver, $G' = (Y, Z, L_Y, L_Z)$. Concatenating the labels (e.g. chromosome identifiers) for all genes in A_i constructs the vertex label function, L_Y. Note that the universal set of vertex labels (e.g. the union of all vertex labels in the canonical quiver) is the union of all labels in a set of genomes and the label of each vertex is therefore a subset of the universal set of vertex labels. Implicitly, we also construct the edge-morphism, $M_V : e \longrightarrow z \mid e_t, e_h \in \bigcup_{i=1}^{n} V_i, \forall z \in Z$, where the tail and head of an edge, e_t and e_h, are a vertices from one of the N genomes in the database. Conceptually, we are merging all edges whose head and tail are part of the same syntenic anchor. Similarly, the concatenation of all edge labels defined by the edge-morphism similarly leads to the construction of the edge-label function, L_Z.

In our implementation, we output the canonical quiver in a GFA-formatted file [155]. Each node is represented with the unique identifier assigned during the database or vertex-morphism construction. The path lines describe the original architecture of a sequence (e.g. chromosome) using the node identifiers and, hence, can be used to extract the edge and vertex labels. We additionally add a genome line starting with the identifier "G" describing the set of sequences for each genome. The resulting GFA-formatted file is portable and can be immediately visualized in any GFA-supported graph visualizer such as Bandage [384].

6.2.4 Structural variant calling using quiver representations

Structural variants are traditionally based on a reference genome, but can also be described as a family of subgraphs each describing architectural similarities and differences across a population. Recall our working example of genomes G_1 and G_2 (see Figure 6.1). We can describe the structural variant as an insertion of three genes in G_2 with respects to G_1. Conversely, we can describe it as a deletion of three genes in G_1 with respects to G_2. In either case, this approach makes use of a reference-genome. However, we can also partition the canonical quiver and describe the graph as a family of subgraphs describing genomic similarities and differences as a population. For example, genes 1, 2 and 3 and genes 4 and 5 can form two disconnected components each describing common genomic architectures between G_1 and G_2. Genes 6, 7, and 8 can also form a disconnected component but instead describe a variant in the genomic architecture between G_1 and G_2.

We identify structural variants in the canonical quiver using a two-step hybrid, reference and population-based approach. We use an inductive graph data structure [385] for representing a canonical quiver enabling us to use a functional paradigm for identifying structural variants. Given a canonical quiver, $G' = (Y, Z, L_Y, L_Z)$, we first define a reference architecture used to partition the quiver into a family set of subgraphs representing differences across the given collection of genomes with respect to a commonly observed population. By default, the reference architecture is obtained by computing the most common genome architecture in the canonical quiver based on the frequency of subpopulations within all edges. Specifically, for a given connected component, we obtain the label of all edges, count the number of occurrences for a given group of labels, and use the label with highest count; resulting in the most co-occurring group of genomes in the canonical quiver—similarly to obtaining the "most weighted path". Optionally, the reference genome architecture can be computed using co-occurrences of sub-populations within nodes rather than edges. For more specific comparisons—for example, comparing pathogenic to non-pathogenic genomes—users can specify a specific population as the reference architecture.

Given the label of the reference architecture, Σ_R, we perform a *reference-cut* operation: we remove all edges, satisfying, $\Sigma_R \subseteq L_E(z) \mid\in \bigcup_{i=1}^{X} \Sigma_i, z \in Z$, followed by the removal of all vertices satisfying, $deg^-(x) = deg^+ = 0$. Conceptually, the reference-cut operation removes edges that are part of the reference architecture followed by nodes with no in or out-edges, indicating genes shared across all genomes. The result is a family of subgraphs, $\Gamma_f \mid f \in F$, where F is the total number of subgraphs, each representing a structural variant with respect to the reference architecture.

Each subgraph Γ_f may contain additional nested structural variation that can be characterized through a recursive labeled-traversal approach. As previously discussed, nested structural variation is generally missed when solely comparing against a reference genome. To characterize the nested variation, we traverse through each Γ_f based on *maximally labeled path traversals*: given some starting node, y_1, and a label, Σ_x we perform a depth-first search traversal to obtain the maximally labeled path, $t = (y_1, y_2, \cdots, y_p$, such that $(y_i, y_j) = (z_i, z_j) \wedge \Sigma_x \subseteq L_E(z) \mid y_i, y_j \in Y, z \in Z, 1 \leq i < j \leq p$. If we remove all edges inferred in t and subsequently rmove all nodes with no in or out-edges, we recreate the reference-cut operation. A recursion-based implementation where a new label is used in

each iteration enables us to dynamically choose a new reference architecture based on all structural similarities and differences within the population of genomes Γ_f.

In our implementation, we first identify all connected components in the canonical quiver. Then for each connected component, we compute the reference architecture, perform the reference-cut operation and tail-recursively report the maximally labeled traversals. Similarly, we store the output in a GFA file where each connected component has a corresponding GFA file describing all family of subgraphs identified and each *path line* describes a maximally labeled traversal.

6.2.5 Ptolemy implementation

All the algorithms discussed are packaged under *Ptolemy* and generalized in three modules. The *extraction + repeat finder* module ($E + R$) creates a database for a given set of genomes and attempts to identify repeat expansions. The *syntenic anchor* module *SA* performs pairwise gene alignments across all genomes in the database, obtain BRHs, and computes syntenic anchors. The *canonical quiver* module (CQ) constructs the canonical quiver by inferring the graph morphism functions from the computed syntenic anchors.

Ptolemy is implemented under a functional paradigm using *Scala* (`https://www.scala-lang.org/`) and released as an open-source software under the *GNU GPL3* license. Binaries, source code, documentation and example datasets are available through *GitHub*: `https://github.com/AbeelLab/ptolemy`.

6.2.6 Benchmark data

We evaluated *Ptolemy* by aligning three different datasets representing various microbial genome architectures and populations. The *MTBC* dataset contains 24 complete assemblies from the *Mycobacterium tuberculosis* complex [190]. The *Yeast* dataset contains 12 complete, PacBio assemblies from the *Saccharomyces sensu strictu* group—the architectures of these genomes were previously analyzed [260]. The *Eco+Shig* dataset contains a mixture of 20 *Escherichia coli* and *Shigella* species that are both commensal and pathogenic—the pan-genome of these organisms were previously analyzed [192]. The accession codes for all assemblies can be found in Supplemental table 1 in [213]. Clustering of assemblies via kmer-profiles was performed with *MASH* [177] using kmer size of 21 and sketch size of 1,000,000. The canonical quivers were visualized using *Bandage* [384] and internal scripts using *Scala*.

6.3 Results

Genome architectures in the microbial world can be diverse rang-ing from species with high sequence conservation to those with only 11% overlap in their genetic content [190, 192]. We therefore evaluated the utility of *Ptolemy* on three microbial datasets representing the spectrum of microbial genetic diversity. The *MTBC* dataset contains complete assemblies from *M. tuberculosis* (22), *M. canetti* (1), and *M. africanum* (1) whose genome architectures are conserved harbouring little structural variation relative to other prokaryotic organisms [190, 191, 386]. The Yeast dataset contains complete assemblies of *S. cerevisiae* (7) and *S. paradoxus* (5) which share large fraction of their synteny but are known to harbour various balanced and in-balanced complex structural variation as well eukaryotic

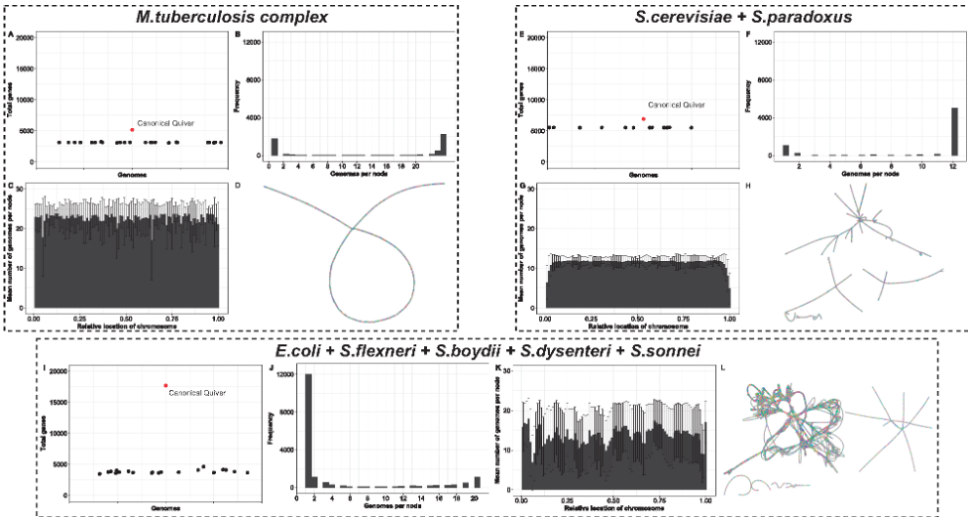


Figure 6.3: Pan-genome and canonical quiver overview of three datasets using *Ptolemy*. The various figures shows an overview of the pan-genome and canonical quiver derived by Ptolemy for M. tuberculosis genomes (top-left), *Saccharomyces* genomes (top-right), and *E. coli* and *Shigella* genomes (bottom). In general, Figures A,E, and I compares the total number of genomes in the canonical quiver in comparison to all genomes in the dataset. Figures B,F, and J shows the distribution of the number of genes shared across all genomes in the dataset. Figures C, G, and K summarizes the Figures B, F, and J as a function of the relative location of the chromosome. Finally, Figures D, H, and L shows a visual representation of the canonical quivers.

horizontal gene transfers [260]. The most diverse set is the *Eco+Shig* dataset consisting of complete assemblies from *E. coli* (13), *S. flexneri* (3), *S. boydii* (2), *S. dysenteriae* (1), and *S. sonnei* (1) which have dynamic genome architectures with many com-plex structural variations and little overlap in their gene content [192]. We inspired our evaluation on previously published analyses of the structural variants and pan-genome—shared fraction of gene content across all genomes—of these datasets [191, 192, 260, 386].

6.3.1 Conserved genome architectures in *MTBC*

MTBC dataset—termed the *canonical quiver*—reflect previously published analyses of the pan-genomes for *Mycobacterium species*. Figures 6.2A-D gives overview summary of the pan-genome derived from the canonical quiver. On average, there are 1,013 more genes in the canonical quiver in comparison to the gene content of the 24 assemblies (see Figure 6.3A)—note that we merge overlapping reading frames into a single, maximal gene (see Methods). Most of these genes are shared across all genomes as 76% of all genes are shared by at least half of the assemblies in the dataset (see Figure 6.3B). In terms of chromosomal locations, we find that the number of genomes containing a gene is constant across the chromosome with no clear "hot-spots" of unique gene content (see Figure 6.3C).

Structural variation encoded in the canonical quivers also reflect previous analyses regarding structural variation within the *MTBC* dataset. Figure 6.2D visualizes the canonical quivers and is (visually) representative of how dynamic the genomes are. As shown, the canonical quiver is largely linear with a single, topological "loop" in the middle. By

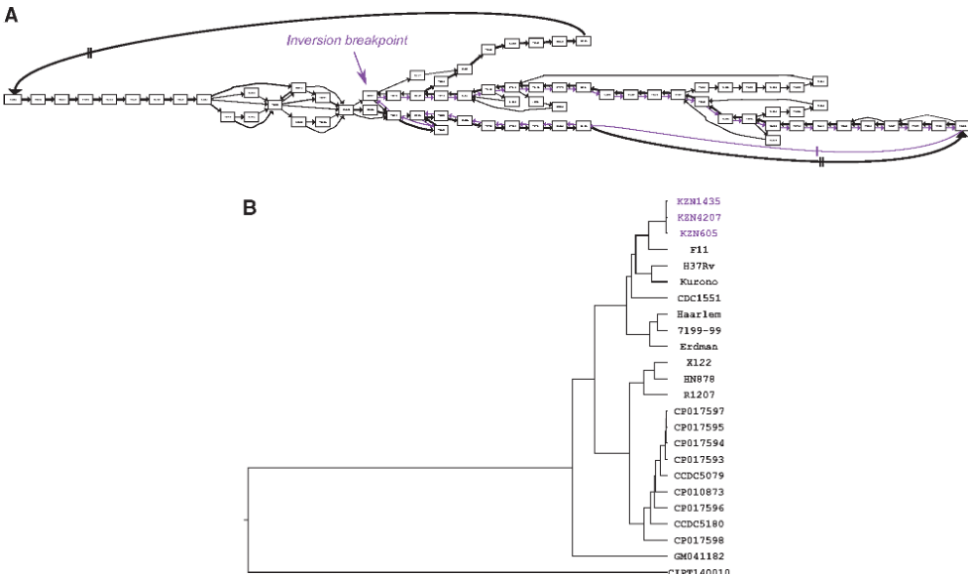


Figure 6.4: Large-scale inversion within a sub-population of *M. tuberculosis* genomes. (A) Shows a subgraph of the canonical quiver at the breakpoint of an inversion present in 3 genomes. Nodes are genes and the edges describe alternative paths that different genomes take: edges are coloured purple when they exclusively describe the three genomes harbouring the inversion, and black otherwise. The thickness of the edge corresponds to the number of genomes traversing the paths—the more common the path the thicker the edge. (B) A dendogram of the hierarchical clustering of all genomes in the dataset based on kmer-profiles. The samples in purples are those harbouring the large-scale inversion which cluster together.

6

extracting the family of subgraphs which correspond to the structural variations in the canonical quiver, we find that the loop is representative of a large-scale inversion in 3 of the 24 genomes (see Figure 6.4A). Kmer-based clustering of the assemblies (see Methods) shows that the genomes harbouring the inversion also cluster together, indicative of a sub-population within this dataset (see Figure 6.4B).

6.3.2 Variable genome architectures in *Yeast*

The canonical quiver confirms previous reports regarding genome architectures in the *Yeast* dataset. Figures 6.3E-H shows an overview of the pan-gnome obtained by *Ptolemy*. The canonical quiver has 6,919 genes, which is on average 1,249 more genes in comparison to the 12 assemblies in the dataset (see Figure 6.3E). Most of the genes in this dataset are universally shared as 80% of the gene content is present in at least half of the assemblies in the dataset (see Figure 6.3F). As shown in Figure 6.2G, the number of genomes per gene is fairly consistent across all chromosomes except for the starting/ending sub-regions where this number sharply falls (see Figure 6.3G).

We were able to identify previously reported structural variation as well as additional variation likely missed due to bias in reference-based comparisons. Although linearity (e.g. synteny) is still observed throughout the quiver, Figure 6.3H shows various topological features reflecting several translocations and inversions. (Note the different connected

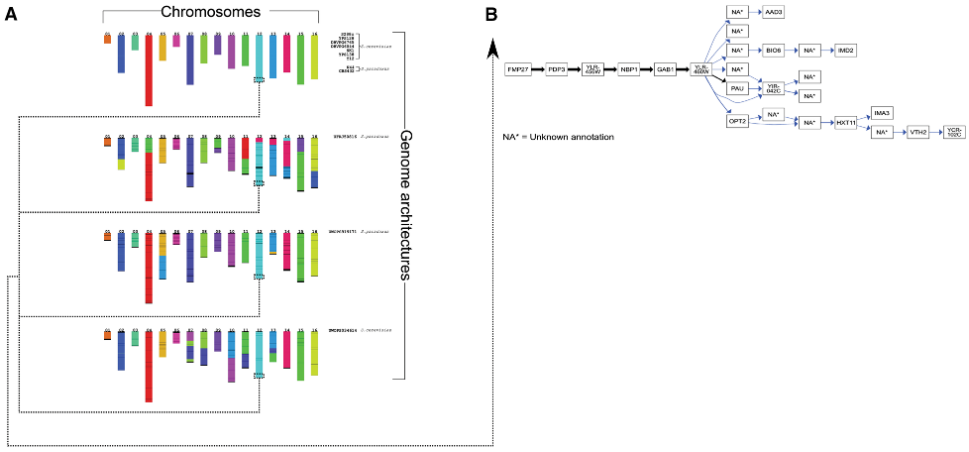


Figure 6.5: Genome-wide and sub-region-specific quiver decomposition for 12 *Saccharomyces* assemblies. (A) shows that the decomposition of the canonical quiver results in 5 unique genome architectures. The first genome architecture (top-most set of chromosomes) is the most common and is largely similar to the commonly used reference genome for *Saccharomyces cerevisiae*, S288C. The remaining three are much more diverse containing several translocation and inver-sions across the 16 chromosomes in the genome. (B) shows a sub-region in the canonical quiver corresponding to the right sub-telomere region of chromosome XII. Black edges correspond to paths containing the reference, S288C, and blue otherwise. Note the additional structural variants present in several genomes which are absent in the reference.

components reflecting different chromosomal sequences in these organisms). By decomposing the quiver, we can reconstruct the genome architectures of the twelve genomes proposed by Yue *et al.* (see Figure 6.5A) [260]. Specifically, the genome architectures for eight genomes are similar to that of the S288C, a commonly used reference genome for *S. cerevisiae* (see Figure 6.5A). For the additional three genomes we find various translocations in inversions across the 16 different chromosomes (see Figure 6.5A).

An example of the type of complex structural variation that exists within the Yeast dataset is shown in Figure 6.5B. The figure corresponds to a sub-graph of the canonical quiver corresponding to the alignment of the right sub-telomere region of chromosome XII. As depicted, there are several structural variants unique to sub-populations in the dataset which are absent in the commonly used reference genome of S288C (see Figure 6.5B). The bottom-most alternative path contains several genes associated to sugar and alcohol metabolism (see Figure 6.5B). These genes are not only unique to 2 of 12 genomes but also contain nested structural variation which is generally missed by reference-based comparisons. An additional example is shown in Figure S1B in [213] depicting the alignment of the right end of the sub-telomere region for chromosome VII. Yue *et al.* previously reported a tandem expansion of two paralogs, MAL31 and MAL33 (involved in the metabolism of the maltose sugar compound), for the *S. paradoxus* genome, CBS432 [260]. We find that this expansion is present—in variable length—in 9 of the 12 genomes and absent only in the *S. cerevisiae* genomes of SK1 and DBVPG6044 along with the commonly used reference, S288C (see Figure S1B in [213]).

6.3.3 A genomic "melting-pot" in the *Eco+Shig* dataset

We observe large variations in the pan-genomes for the 20 assemblies in the *Eco+Shig* dataset. Each genome contains about 3,825 genes, contrasted by the canonical quiver which has a total of 17,698 genes (see Figure 6.3I). This variation is further highlighted in Figure 6.3J where only 18% of all genes are shared by at least half of the assemblies in the dataset. Furthermore, the number of genomes per gene is highly variable and varies throughout the chromosome (see Figure 6.3K).

We investigated structural variation encoded in the canonical quiver by comparing the genome architectures of commensal and non-commensal pathogens [192]. The complex structure of the canonical quiver is shown in Figure 3L and highlights the dramatic variation that exists within the genomes of the *Eco+Shig* dataset. Although some linearity exists, Figure 6.3L shows that the canonical quiver contains many complex topological features representing various forms of structural variations, inversions, and horizontal gene transfers. (Note that a subset of these genomes contain several plasmid sequences and, hence, Figure 6.3L displays several connected components). In the *Eco+Shig* dataset, 9 genomes are described as commensal while the remaining 11 genomes are described as pathogenic [192]. We defined the reference genome architecture to the 9 commensal genomes (see Methods) and extracted the family of subgraphs representing structural variation between the two populations.

We found 50 structural variants exclusive to the pathogenic genomes of containing at least three genes and shared by at least two genomes. Among the largest structural variant is a sub-graph in the canonical quiver of about ~24 genes in length that is exclusive to four *Shigella* genomes: S. flexneri strains 2a 301 and BS12, S. dysenteriae strain Sd197, and S. sonnei strain Ss046 (see Figure S2 in [213]). Closer analysis showed that this variant corresponds to the virulence-based type III secretion system [387], a hallmark genetic compo-nent in pathogenic bacterial species [388].

6.3.4 Performance of *Ptolemy*

Although the construction of the canonical quiver can be fast—e.g. ~10 min for 24 genomes (see Table 6.1)—it's important to note that the time complexity is ultimately $O(n^2)$. The two most computationally heavy steps in *Ptolemy* is computing best reciprocal hits (BRHs)—which currently uses pairwise gene alignments across all pairs of genomes—and the syntenic scoring of each BRH, each which is $O(n^2)$ (see Table 6.1). For the latter step, the worst case scenario is comparing highly conserved genomes (such as *Mycobacterium tuberculosis* as done in this study). For this type of organisms, many genes are shared across a large fraction of all genomes and nearly every gene will have a BRH across all genomes, resulting in n^2 number of synteny scorings. Given that *Ptolemy* is implemented under a functional paradigm and nearly entirely immutable, these steps are easily parallelizable and currently makes use of all available CPUs. Analyzing large data sets is, in part, dependent on the number of available CPUs in a machine/cluster. As an example, we ran *Ptolemy* on 100 *Mycobacterium tuberculosis* genomes which took a total of 1 hour and 32 minutes using 20 CPUs.

Table 6.1: Run time of *Ptolemy* across three datasets. *Ptolemy* is separated in three modules: *extraction + repeat finder* ($E+R$), *syntenic anchors*, (SA), and construction of the *canonical quiver* (CQ).

Dataset	Genomes	Module	Wall clock (min:s)	Max mem. (Gb)	CPUs
MTBC	24				
		$E+R$	0:43	0.680	1
		SA	11:35	1.35	4
		C	0:03	-	1
Yeast	12				
		$E+R$	0:45	0.694	1
		SA	5:05	1.44	4
		C	0:03	-	1
Eco+Shig	20				
		$E+R$	0:42	0.632	1
		SA	4:50	1.33	4
		C	0:03	-	1

6.4 Discussion

Advances in long-read sequencing technology are enabling re-searchers to feasibly acquire "complete" assemblies for a collection of microbes. As this technology becomes more accessible, we can begin to shed light at the diversity of genome architectures across different (sub-) populations of microbial species, which has largely been hindered by limitations of reference-based computational approaches. In this paper, we present *Ptolemy*: a reference-free method for analyzing genome architectures across a collection of microbial genomes. *Ptolemy* represents each genome as a labelled-multi-directed graph, known as *quivers*. Using synteny analysis, the quivers can be merged into a single, canonical quiver representing a structural-based multiple whole genome alignment. As shown in the application of *Ptolemy* across three different datasets of *Mycobacterium*, *Saccharomyces*, and *Escherichia* and *Shigella* species, the canonical quiver can be used to study pan-genomes as well as systematically discovering structural variants in context of (sub-)populations.

The application of *Ptolemy* on the three dataset shows the spectrum of genomic diversity that can exists in the microbial world. For example, the pan-genomes of the *MTBC* dataset confirm high conservation of the genome architectures of these organisms, which harbour rela-tively little-structural variation [190, 191]. Structural variants in these organisms are therefore used as lineage-specific markers [365, 389]. Specifically, we show that traversals of the canonical quiver can identify a large-scale inversion that exists within 3 of the 24 genomes (see Figure 6.3); these genomes correspond to a family of highly virulent strains endemic to a sub-region in South Africa where the inversion has been previously observed [386]. It is important to note that Figure 6.3B shows roughly 2,000 unique genes across the 24 assemblies. Closer analysis showed that the majority of these genes correspond to transposable insertion sequences and PE/PPE genes which are repetitive and variable across genomes [390–392]—the latter which correspond to ~10% of gene content in *M. tuberculosis* genomes [390, 391].

For *Saccharomyces* species, sub-telomeric regions—the first/last ~20-30 Kbp of a chromosome—are biologically relevant as they harbour gene families that heavily influence biotechnology-

based phenotypes [207, 234, 260]. However, these regions are notoriously challenging to compare across different genomes as they typically undergo gene-deletion, expansion, and reshuffling leading to highly dynamic architectures [207, 234, 260]. Indeed, Figure 6.3G re-confirms previous observations of the diversity in these regions showing that the genes in the beginning/end of each chromosomes are not commonly found across all genomes. More specifically, Figure 6.4B shows the alignment of the right sub-telomeric region of chromosome XII across all genomes highlighting nested-structural variation unique to sub-populations in the dataset.

Expectedly, the results obtained in the *Eco+Shig* dataset dramatically differs to those of the *MTBC* and *Yeast* dataset. We observe a significant lower number of genes shared across all genomes similar to those previously reported (see Figure 6.3 I and J) and find more complex structural variation in the canonical quivers (see Figures 6.2 D, H, and L). Such dynamic genome architectures can complicate comparative studies [192, 368]. Our ability to identify structural variation—specifically between commensal and pathogenic strains—highlights the viability of *Ptolemy* in different microbial populations.

The accuracy of the *Ptolemy* is depended on the accuracy of the gene annotations in a given dataset. *Ptolemy* only compares the sequence within the boundaries of each gene and is therefore sensitive to annotations errors. More specifically, annotation errors can lead to false negative merging of nodes inducing false positive structural variants. This is shown in Figure S1A in [213] where the upper-most path of the alignment in the right sub-telomeric region of chromosome V is likely caused by gene annotations errors: the sum of the size of the two adjajcent TOG1 annotations is approximately the same as the size of the TOG1 annotation in the bottom, adjacent path. Therefore, the alternative path will be identified as a structural variant although it is likely that this is the same sequence present in the remaining genomes in the dataset (see Figure S1A in [213]). We acknowledge that annotating genomes is an error-prone process and often requires manual curations [393, 394]. For this reason, the current implementation of *Ptolemy* is regarded as an "approximate structural aligner" and care should be taken when comparing genomes of un-known annotation quality. Nevertheless, we were still able to construct pan-genomes and identify structural variants that agree with previous published studies despite using genomes sequenced and annotated by different groups and pipelines (see Figure 6.3, 6.4, and 6.5).

Future work could use a two-step alignment process: syntenic-anchoring followed by local-realignments of nodes. This is primarily to refine alignments of repetitive sequences, especially those involved/nearby repeat expansions. As discussed in the Results, Figure S1B in [213] shows a sub-graph of the canonical quiver of *Yeast* dataset repre-senting the alignment of right sub-telomeric region of chromosome VII. We show that there is a tandem expansion of variable length for two paralogous genes across 9 of the 12 genomes. In this alignment, the right-flanking genes, PAU, COS2, and COS6, are present in other sub-populations and are considered BRHs but *Ptolemy* considers them unique for most of the genomes. This is largely due to difficulties in scoring the synteny in the surrounding region heavily influenced by the downstream repeat expansion as well as the different genes present upstream in each genome. Therefore, a two-step approach may first build the canonical quiver and followed by a traversal seeking to re-score the synteny of genes that are considered unique but possess BRHs in some defined neighbourhood of a local

subgraph.

6.5 Conclusion

Advances in sequencing technology is expanding our knowledge of the genetic diversity in microbial populations. Lacking are computational methods that can simultaneously compare multiple assemblies without restricting analysis to only "similar" genomes. Here, we show that *Ptolemy* is a flexible method that can systematically identify structural variation across a collection of assemblies while providing insights in the population structure and pan-genome of the collection—all without the need of a reference. *Ptolemy* tackles long-standing challenges in comparative genomics including independence from a reference genome, characterization of complex structural variation as sub-populations, and viability in studying both conserved and highly dynamic genomes. The work presented here is a step forward for studying the genetic diversity that is yet to be characterized in the microbial world.

7

An educational guide for nanopore sequencing in the classroom

The last decade has witnessed a remarkable increase in our ability to measure genetic information. Advancements of sequencing technologies are challenging the existing methods of data storage and analysis. While methods to cope with the data deluge are progressing, many biologists have lagged behind due to the fast pace of computational advancements and tools available to address their scientific questions. Future generations of biologists must be more computationally aware and capable. This means they should be trained to give them the computational skills to keep pace with technological developments. Here, we propose a model that bridges experimental and bioinformatics concepts using the Oxford Nanopore Technologies (ONT) sequencing platform. We provide both a guide to begin to empower the new generation of educators, scientists, and students in performing long-read assembly of bacterial and bacteriophage genomes and a standalone virtual machine containing all the required software and learning materials for the course.

7.1 Introduction

What defines a biologist? In short, a biologist is a person who studies life and living organisms. But this simple definition hides the true complexity of the field of biology. Biology covers diverse topics such as molecular biology, structural biology, ecology, evolution, genetics, microbiology, immunology, and biotechnology. Importantly, most (if not all) of these topics have undergone incredible progress due to rapid discoveries and technological advances [395, 396]. As such, a modern biologist has the inevitable tasks of adapting to rapid change and mastering new knowledge and technology.

One of the most important revolutions in the field of biology was caused by the development of next-generation sequencing (NGS) technologies. Using massively parallel processing of samples, NGS dramatically reduces sequencing time and costs, enabling the sequencing of entire genomes. Currently, genome sequencing and analysis have become a crucial component in biology, as evidenced by recent scientific breakthroughs [81, 397]

This chapter have been published in *PLOS Computational Biology* [214]

and by the exponential increase of reported genomes on GenBank (e.g., from 30,000 sequenced prokaryotic genomes in 2014 [398] to 183,000 in 2018 (https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/), a 6-fold increase in only 4 years). Thus, not only do biologists need to adapt and learn how to use these emerging technologies, they also need to learn how to mine the ever-growing mountain of genomic information they generate, which requires bioinformatics skills. Now, the question is how do we train this generation of biologists so that they have the required computational skills?

7.2 Bridging bioinformatics to biologists

Over the past few years, we have taught introductory bioinformatics to undergraduate (second year BSc) biology students with basic molecular biology training. They are versed in standard techniques (such as basic DNA extractions and PCR) but are unfamiliar with specific DNA sequencing chemistries. In the past, this mandatory computational course was entirely disconnected from lab work, making it hard for students to grasp how bioinformatics and biology are connected. To address this disconnect, we here share a more integrated approach to teach bioinformatics to biology students. These students have a conceptual grasp of sequencing and bioinformatics but not the detailed view on how various lab techniques (e.g., NGS chemistries) combined with various analysis methods (e.g., assembly, variant calling) can be used to answer specific biological questions and how these techniques interact with each other.

The overall idea is to start from where students are already familiar (i.e., biology) and expand from there. There are 4 types of learning activities in the course (see Figure 7.1): (1) lectures in which students receive classroom instruction on bioinformatics topics, (2) practical sessions in which students apply the material from the lectures to solve practical exercises supervised by teaching assistants, (3) lab work in which sequencing data are generated, and (4) a project that applies the bioinformatics concepts learned in the lectures on data from the lab work. This is concluded by a poster session in which all students get to review each other's work. A week by week overview can be found in S1 Table in [214].

The formula presented here focuses on introducing bioinformatics to biology students, helping them to acquire the skills and insights needed to operate and troubleshoot existing algorithms. The course does not focus on developing skills needed to create novel algorithms or models.

During the pilot run of this course in the academic year from 2017 to 2018, we used Oxford Nanopore Technologies (ONT) MinION sequencing as a data generation platform. This platform was selected because it has low capital cost and is a new exciting technology easy to engage students with. Real-time data acquisition gives immediate feedback to the students that data are being produced, even if they have to keep it running overnight. It is easy to imagine they could get one of these devices at home. Students can see themselves as scientists, as people discovering something new, an idea that we really like to foster. Ultimately, any fast, cheap, and accessible sequencing platform would be good for our goals, yet only MinION is currently available.

MinION has already made its way into undergraduate and graduate courses [399, 400]. Some of these courses focused on data analysis; they organized hackathons in which students needed to devise a pipeline to infer the ingredients of food DNA samples or identify human DNA samples [399]. Others developed the application of MinION further by also

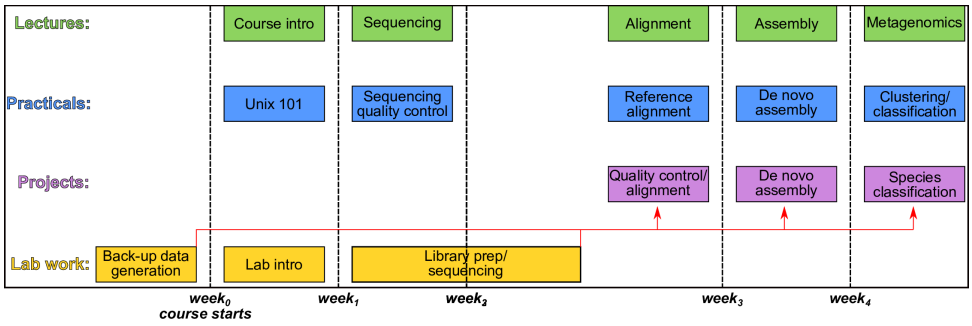


Figure 7.1: Integrated bioinformatics training with time on the x-axis. Lectures (green) give students the necessary background to execute and understand Practical (blue) and Project (purple) sessions. Laboratory sessions (yellow) enable students to employ their biological background and prepare their own DNA libraries from samples of interest. Libraries prepared by each student group are pooled together and run on a MinION device (Oxford Nanopore Technologies, Oxford, UK), generating data to be processed in Project sessions. Backup data previously prepared from the same samples can be used if the students' MinION run fails to provide enough quality data for analysis. In the Practical sessions, students learn to use established bioinformatics methods, with an emphasis on processing long-read data (see Figure 7.2, S1 Table and S1 Text in [214]). In the Project sessions, they then apply these methods to the generated data to answer specific research questions. After intragroup and intergroup discussions of results, students prepare their final project report and present their results in a poster format.

teaching laboratory techniques for DNA extraction and sequencing library preparation [400].

Additionally, the portable size of ONT's MinION and the simplicity of library preparation enable scientists to use this technology in a wide variety of environments, including a standard classroom [401–403]. As such, this device is not only attractive for researchers but also for educational instructors: If this technology is empowering scientists to embark on novel scientific studies, why not also empower young students to embark on effective educational experiences?

7.3 Integrating nanopore sequencing in the classroom

The challenge set for students in our course was to identify and discover novel phages from environmental samples and to reconstruct complete genomes from single-isolate and metagenomics samples. The students had to address the following research questions, which were introduced at the very beginning of the course: (1) Can we assemble and annotate fully closed genomes from a small number of long reads? (2) What are the considerations for the assembly of metagenomics samples compared to single isolates? (3) What is the advantage of long-read sequencing for the analysis of metagenomics samples? (4) Can we identify virulent and temperate phages in metagenomics samples? (5) What genes of interest can we find in both bacteria and phage genomes? Twenty-four groups of 4 students (96 total) prepared their own DNA libraries of various single-isolate bacterial, bacteriophage, and metagenomic samples in the classroom. Number of groups and their size were determined to allow for sufficient supervision within the available lab space. If possible, smaller groups are preferable to increase the hands-on time of each student. We would like to emphasize the benefits of having multiple groups working on different

related samples (e.g., each barcode represents a similar but different microbial isolate). This allows groups to initiate discussions about differences in their own findings—such as unique sequences, structural variants and presence and/or absence of genes—and hypothesize how those differences may influence the phenotypic traits of their sample. This exercise helps them further appreciate the value of bioinformatics skills in a biological setting and how the two are ultimately connected.

The DNA libraries were prepared using the rapid barcoding kit (SQK-RBK004), which has fewer steps than other available kits and thus allows the procedure to be completed within the 3-hour timeframe of the class. For longer sessions, the ligation sequencing kit (SQK-LSK109) could be used, increasing the robustness and throughput of the experiment. Both kits allow for barcoding of multiple genomic DNA samples. Samples were prepared individually by each group and then barcoded and pooled together at different proportions depending on the success of each group. When sequencing runs failed, the student was supplied with previously generated backup data.

After running DNA samples in MinION, students performed quality control of their data and then assembled the genomes. As we focused on teaching technical concepts of bioinformatics, we provided a computational guide (see S1 Text in [214] and summary in Fig 7.2) containing ready-to-go commands and scripts for commonly performed tasks that can be broadly used with MinION data. To facilitate the use of this guide, we provided a standalone virtual machine containing all required software used in S1 Text in [214].

Once data processing was completed, students pursued a variety of research questions, such as investigating the genomic composition of their bacterial sample as well as the population composition of their metagenomics sample. For example, students would determine the bacteriophage species in their barcoded sample and compare their assembled genome to that of the closest reference genome found in the National Center for Biotechnology Information (NCBI) reference sequencing database (RefSeq). In all cases, students found that their assembly had little overlap with the reference, prompting discussions about the novelty of the genetic content in their phage.

Students ran Centrifuge [408], a species classification and quantification tool, on their metagenomics sample and generally concluded a mixture of viral and bacterial species. This process stimulated discussion about a number of course-related topics: (1) limitations of kmer-based tools (e.g., kmers are not always unique to individual species), (2) biases when comparing against a reference data set (e.g., you can only classify what you have previously observed), (3) understanding bacteriophage biology (e.g., phages can integrate their DNA in a bacterial host; therefore, sequences that are labeled as "bacteria" may actually correspond to integrated phage DNA), and (4) understanding whether long-read sequencing is advantageous to the scientific question addressed (e.g., long-read sequencing helps improve assembly quality of metagenomes, but the high error rates of the technology still limit its usefulness; here, combining short-read and long-read data could be the best approach to improved contiguity and base pair–level accuracy). These topics were framed to explore how they may affect the student's computational observations.

Through the integrated approach in our course, students can easily grasp the direct influence of the experimental protocol on data quality. For example, a student's excessive pipetting leads to observably shorter read-length distributions, resulting in fewer unique overlaps in the pairwise alignments, a less contiguous assembly graph, and ultimately

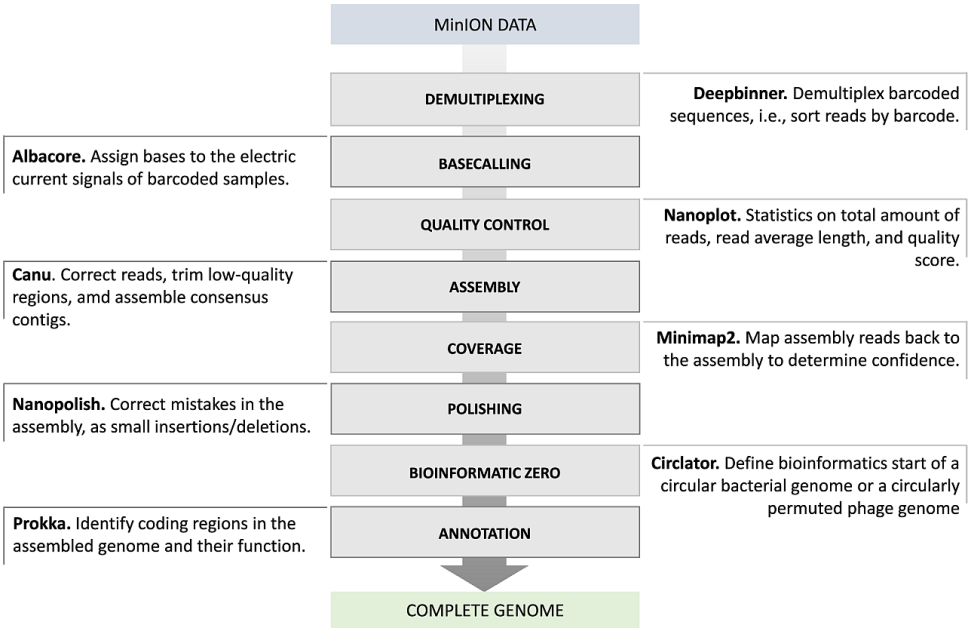


Figure 7.2: Pipeline for genome assembly using MinION data. First, the barcoded sequences are demultiplexed using *Deepbinner* [404] and basecalled using *Albacore* (Oxford Nanopore Technologies, Oxford, UK). *Nanoplot* [405] is used to assess the quality of the sequencing data for downstream processing. If the data have sufficient quality, they are used for assembly using, e.g., *Canu* [151]. Confidence on the resulting consensus assembly is obtained using *Minimap2* [93]. The assembly is polished to remove common mistakes using *Nanopolish* [15], and then Circlator [406] is used to determine the zero-based start of the genome, which depends on whether it is a bacterial sequence or a bacteriophage sequence. Finally, the assembled genome is annotated using *Prokka* [407]. Please refer to S1 Text in [214] for further details.

more fragmented assemblies. Furthermore, the setup is sufficiently generic that different scientific questions could be addressed using this pipeline, and it is sufficiently flexible to adjust to the students' background.

We experienced increased interest and engagement in our course from both the instructors and the students. Students were much more interested in the course content because they could assume scientific responsibility and ownership. Spending several hours or days in the lab goes a long way to make "scientists-to-be" feel "this is my data."

The instructors leveraged the practical classes as an opportunity to generate and analyze data for potential pilot studies, i.e., preliminary data for the next round of grants. In our pilot version of the course, the experiments were chosen such that they contribute to ongoing research in the lab. As a result, we generated several follow-up project ideas, one of which resulted in a master's thesis on heterogeneity of bacteriophage genomes detected by nanopore sequencing, as well as a tripling of the number of undergraduate lab-rotations in the area of bioinformatics.

Naturally, many of the assignments, including interpretation and comparison of a genome assembly from single bacterial isolates to that of viral samples, were open-ended and initially challenged the students. However, the experience gave them a more realistic impression of academic research and foundational skills to help them in their future career as modern biologists. In particular, different samples required different data interpretations, naturally spurring discussions and collaborations among students. Future editions of such an integrated course could consider even developing the student ownership further by explaining the "problem" and asking students to design the DNA sequencing experiments given the boundaries of the reagents available. With adequate supervision and coaching to include proper controls and experiments, this could lead to even greater collaboration and ownership by the students.

7.4 Conclusion

Considering the fast pace at which sequencing technologies progress and at which genomics data are generated, it is no longer possible to ignore the urgency of equipping young biologists with the required skills to manage the amount and type of sequencing data being generated. Here, we used nanopore sequencing as one possible tool to prepare a new generation of bioinformatics-aware modern biologists. Nanopore sequencing offers an exciting opportunity to not only introduce students to the field of genomics and bioinformatics but also to address advanced biological and computational problems. Simple customizations of the assignments are possible to make the course different every year and to make it suitable for teaching students of different backgrounds, such as computer science (e.g., toolbox handling, algorithm understanding), molecular biology (e.g., genomics, sequencing), or medicine (e.g., pathogen detection, cancer diagnostics). MinION also gives a chance to teach the students how to use different tools and community-based analysis and the importance of constantly updating their knowledge of recent technological developments.

The virtual machine and guide provided herein intend to assist science educators and also geneticists to address timely questions in biology, such as detection of epigenetic modifications, characterization of human genetic variation, real-time detection of pathogens, characterization of structural variation in cancer, and analysis of population transcrip-

tomics. A walkthrough of ONTassembly of prokaryotic genomes and their viruses is provided in S1 Text [214]. All materials, including the virtual machine image, are available at `https://github.com/AbeelLab/integrated_bioinformatics`.

8

Discussion

With the rapid progression of genome sequencing technology, microbiology has embraced bioinformatics as a core component of its research. Indeed, the chapters in this thesis showcase the power of (novel) computational methods along with recent sequencing technologies as they have unraveled biological insights about the genomes of influential microbial organisms, particularly in yeast. However, reflecting the 60+ years of the bioinformatics research in conjunction with recent scientific advancements, there is still a lot left to explore. Here, I share some thoughts on-going challenges in bioinformatics, focusing on the following question: how should one compare n genomes in light of recent technological and algorithmic developments?

Since the sequencing of the very first genes, researchers have postulated the promise of comparative genomics, that is, comparing (multiple) genomes of similar and/or different organisms in order to understand their biology and evolutionary history [75, 185, 186, 188, 189]. But as discussed in the introduction, technological challenges in whole-genome sequencing and assembly has restricted us to either a limited number of sub-regions and/or samples. For example, early whole-genome sequencing data from the 1990s and early 2000s provided complete microbial assemblies, but their high-cost and labor-intensive protocols restricted the total number of genomes that could be affordably sequenced [81, 109]. Nevertheless, ambitions to understand their biology and evolutionary history is reflected in the various comparative methods that have since been developed. This was followed by second-generation sequencing data, which was less costly enabling us to sequence many more samples, but at the trade-off of fragmented and incomplete assemblies [81, 109]. We could still compare genomes deriving from second-generation sequencing data through the (short-read) alignment and variant-calling paradigm, and various methods following that paradigm were indeed developed [108, 409]. However, this approach is knowingly biased by the choice of the reference genome and fails to interrogate "inaccessible" subregions that are repetitive and or difficult to assemble.

Today, third-generation sequencing technology combines the best of both worlds by offering the ability to obtain complete genome assemblies in a high-throughput and affordable manner. Excitingly, we are beginning to routinely obtain high-quality and/or compete genomes, especially for microbial organisms [109, 410]. And so, if the chal-

lenge is not longer purely an issue of assembling genomes, the one can begin to focus in the ”comparative”-part of comparative genomics. But with the availability of complete genome assemblies, this begs the question: how do we better utilize the additional information from complete assemblies? As most methodological questions goes, it depends what we are trying to answer.

8.1 Systematic variant calling from multi-whole genome alignments?

A valuable use of the align-and-variant-call paradigm established with second-generation sequencing is that it enables systematic identifications of variants across many genomes, which we can subsequently associate to phenotypes. Unfortunately, it is much more challenging to systematically identify complex structural variations from only short-reads , and hence, genome-wide associations studies have largely been limited to single-nucleotide polymorphisms. With complete genome assemblies, could we systematically identify the “complete set” of variation without any biases, enabling more enriched associations?

One obvious challenge is properly defining collinear regions. This may be relatively easier when comparing near-clonal populations due to a high abundance of anchors which simplify a multi-genome chaining step, but becomes harder when the sub-population diverge in sequence similarity [365, 411, 412], especially if a minimum number of anchors for a chain is enforced. Of course, one could argue that if a subset of sequences are too divergent, then they should be left-out of the collinear blocks, leading to a set of variants unique to some collection of individual genomes. However, in (microbial) coding regions, variation in DNA sequence does not necessarily imply changes to amino acids [413, 414], so DNA subsequences from some population may actually be conserved at the protein level, and should not necessarily be seen as distinct sequences. This is further reinforced when considering that adaptations to codon usage which can influence the efficiency of expressed proteins, even if their DNA changes are synonymous [413, 414].

Additionally, how do you handle structural variation? For example, there are different approaches to multi-whole-genome alignment, such as positional homology and *glocal* alignments [103]. This choice becomes important if one cares more about structural variation, as opposed to aligning all homologous sub-regions regardless of their locations in the genome. For example, bacterial genomes can undergo large structural changes due to integration and/or rearrangement via horizontally acquired sequences. Similarly, unrelated bacterial genomes can acquire shared genomic sequences that provide them with similar phenotypic behaviour. Alternatively, one could ’revert’ structural variation in all genomes to match that of a single reference genome, enabling a single, artificial large collinear region. However, this does not guarantee a common a collinear architecture, especially if the reference genome fails to contain sequences common across the remaining genomes.

There is also the question of how to perform the alignments of collinear blocks: a progressive approach is fast but can introduces biases by fixing the position of indels despite new information from later alignments [415]. Alternatively, a partial-order alignment graph approach can better represent indels, but the sequences are order dependent [416]. Even then, these are still heuristically based with defined match and gap penalties, should

those parameters dynamically adapt throughout a population of sequences [415]?

There has been a growing interest in genome graphs, aiming to facilitate comparisons of multiple genomes by summarizing and navigating them through a graph-like data structure [174]. In principle, genome-graphs can directly represent multiple-whole genome alignments: at the sequence level, they can natively support partial-order-like alignments of collinear region, or at the very least their multiple-sequence alignment as a sub-graph with different nodes and edges. As such, variants in this representation are not only single allele differences, but also alternate paths (e.g. haplotypes). At the structural level, all collinear regions could be connected by their adjacent locations in their native genomes, enabling one identify structural variants such as inversions, translocations, and large deletions and insertions through traversals of every collinear region. Thus, genome graphs could facilitate comparisons of multiple whole genomes, but the precise methodology on how to construct them remains an open problem.

8.2 The phasing of metagenomes

Metagenomic analysis is a comparison of large collection of microbial genomes. As one would expect, long-read sequencing provides more complete reconstructions of their genomic landscapes. Importantly, it also offers opportunities to understand intra-strain diversity, much in the same manner that long-read provides better opportunities for haplotype phasing of multi-ploid organisms. Differences in alleles in two strains of the same species can lead to different protein sequences (and in some cases different genes), altering their biological capabilities; such as in the case in mix tuberculosis infections. While different genome architectures—such as presence, absence, and arrangement of genes in a genome—can lead to different regulations of biological functions. This is particularly emphasized in a recent study showing that structural variation within microbial communities in the human microbiome may explain risks to metabolic diseases [417]. And although it would obviously vary, the possibility that different microbiomes—say healthy and diseased guts of human individuals—harbour different levels of intra-strain diversity could further aid our understanding in the role of microbiomes and diseases. Nevertheless, most long-read assembly algorithms currently only report a consensus representation of the genomes in the community, ignoring diversity of alleles and genome architectures. But the characterization of intra-strain diversity in metagenomic datasets isn't all too different from haplotype phasing of non-haploid genomes.

One can imagine applying existing frameworks for characterizing heterozygosity, such as a *de novo* assembly approach aiming capture haplotype information [170] or through a two-step approach by first generating a consensus-draft assembly and iteratively aligning reads to detect heterozygous variants. However, the uncertainty of the number of heterozygous genomes (comparable to the problem of chromosome copy number), variability of coverage due to mix-microbial communities and biases in DNA extraction methods, as well as horizontal gene transfers further challenges the formulation of this problem. Nevertheless, there are some (methodological) insights that can still be adopted.

Large heterozygous structural variants of several hundred nucleotides are obvious genomic features to resolve using long-reads. For example, contextual information (e.g. haplotypes) about the surrounding regions of a structural variation would already provide a higher resolution of the intra-strain diversity different microbial populations, such as the

one generated by David Zeevi et al. in 2019 [417]. Although one would expect a logarithmic decrease in the frequency of heterozygous structural variants as their size increases, their origin (such as attributing them to different species or different strains) as well as the genomic structures they impose (such as characterizing instances of horizontal gene transfer and operon structures) is not well explored and can be enriching to association studies. As such, one may think that identifying large bubbles in a metagenomic assembly graph or generating consensus draft assemblies and re-aligning the original long-reads may be suffice to already provide enough contextual information to identify structural haplotypes. But even then, their detection and characterization may not always be straightforward, as inter-species homologous regions and repetitive sequences can complicate the construction of the assembly graph itself leading to detections of false positive and false negative structural variants due to under- or collapsed representations of these sequences. As such, it may be more advantageous to use a method utilizing a combination of both approaches to precisely interrogate such variants.

Importantly, not all heterozygous variants are large. Smaller variation—such as SNPs and indels of a few nucleotides—can alter the function of proteins, are thus equally important to detect. These variants are more challenging to identify due to the nosiy nature of long-reads making it challenging to discern true variation from sequencing errors. In a *de novo* assembly approach, it is a common practice to "merge" sequences that look similar enough to either correct reads or simplify the assembly graph, consequently masking sequence heterozygosity. Furthermore, read-error corrections can also mask heterozygosity, requiring one to refer back to the raw reads. In an alignment approach, variants generally require certain coverage thresholds, which itself is non-uniform due to differences in abundance in mix-microbial communities.

Simple cases are single (or few) variants surrounded by conserved sequences that result in simple bubbles in an assembly graph, or clear alternate alleles in an alignment pileup. And with long-reads, can be easily traversed to yield different bacterial haplotypes. More challenging are stretches of contiguous sequences that have diverged leading to paths constituted by many (nested) bubbles or contiguous stretches of various alternate alleles. These signals could correspond to different evolutionary adaptations of (non-)coding sequences, including divergence in codon usage and protein functionality, but would be challenging to discern from sequencing errors if such regions are under-sampled due to low-abundances of such populations. Under the assumption that sequencing errors are uniformly distributed across long-reads, one could imagine specific situations where only a particular sub-region has diverged. Discerning such cases from sequencing errors could thus potentially be accomplished by comparing paired-distributions of sequence similarity across the reads relative to the assembly graph or linear draft assembly. Ultimately, although there are heuristic approaches to resolve small heterozygous sequences the relatively high error-rates of long-read along with large fluctuations in coverage may not be enough to provide sufficient haplotype information.

Bibliography

References

[1] Kimberley J. Hockings, Nicola Bryson-Morrison, Susana Carvalho, Michiko Fujisawa, Tatyana Humle, William C. McGrew, Miho Nakamura, Gaku Ohashi, Yumi Yamanashi, Gen Yamakoshi, and Tetsuro Matsuzawa. Tools to tipple: ethanol ingestion by wild chimpanzees using leaf-sponges. *Royal Society Open Science*, 2(6):150150, 2015.

[2] Franz G. Meussdoerffer. *A Comprehensive History of Beer Brewing*, chapter 1, pages 1–42. John Wiley & Sons, Ltd, 2009.

[3] Thomas Pfeiffer and Annabel Morley. An evolutionary perspective on the crabtree effect. *Frontiers in Molecular Biosciences*, 1:17, 2014.

[4] Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[5] Adelfo Escalante, David R. López Soto, Judith E. Velázquez Gutiérrez, Martha Giles-Gómez, Francisco Bolívar, and Agustín López-Munguía. Pulque, a traditional Mexican alcoholic fermented beverage: Historical, microbiological, and technical aspects, 2016.

[6] Sofia Dashko, Nerve Zhou, Concetta Compagno, and Jure Piškur. Why, when, and how did yeast evolve alcoholic fermentation?, 2014.

[7] Dara N. Orbach, Nina Veselka, Yvonne Dzal, Louis Lazure, and M. Brock Fenton. Drinking and flying: Does alcohol consumption affect the flight and echolocation performance of phyllostomid bats? *PLoS ONE*, 2010.

[8] Frank Wiens, Annette Zitzmann, Marc André Lachance, Michel Yegles, Fritz Pragst, Friedrich M. Wurst, Dietrich Von Holst, Leng Guan Saw, and Rainer Spanagel. Chronic intake of fermented floral nectar by wild treeshrews. *Proceedings of the National Academy of Sciences of the United States of America*, 2008.

[9] Matthew A. Carrigan, Oleg Uryasev, Carole B. Frye, Blair L. Eckman, Candace R. Myers, Thomas D. Hurley, and Steven A. Benner. Hominids adapted to metabolize ethanol long before human-directed fermentation. *Proceedings of the National Academy of Sciences of the United States of America*, 2015.

[10] Theodore Robert Dudley. *The drunken monkey: why we drink and abuse alcohol*. University of California Press, 1st editio edition, 2014.

[11] Hannah Ritchie. Alcohol consumption. *Our World in Data*, 2018. https://ourworldindata.org/alcohol-consumption.

[12] Glen Fox. Starch in Brewing Applications. In *Starch in Food: Structure, Function and Applications: Second Edition*. 2017.

[13] Nore Struyf, Eva Van der Maelen, Sami Hemdane, Joran Verspreet, Kevin J Verstrepen, and Christophe M Courtin. Bread Dough and Baker's Yeast: An Uplifting Synergy. *Comprehensive Reviews in Food Science and Food Safety*, 16(5):850–867, sep 2017.

[14] D. Wang, Y. Xu, J. Hu, and G. Zhao. Fermentation kinetics of different sugars by apple wine yeast Saccharomyces cerevisiae. *Journal of the Institute of Brewing*, 2004.

[15] Brigida Gallone, Jan Steensels, Troels Prahl, Leah Soriaga, Veerle Saels, Beatriz Herrera-Malaver, Adriaan Merlevede, Miguel Roncoroni, Karin Voordeckers, Loren Miraglia, Clotilde Teiling, Brian Steffy, Maryann Taylor, Ariel Schwartz, Toby Richardson, Christopher White, Guy Baele, Steven Maere, and Kevin J. Verstrepen. Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell*, 2016.

[16] Ofer Bar-Yosef. The natufian culture in the levant, threshold to the origins of agriculture. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):159–177, 1998.

[17] Tobias Richter, Amaia Arranz-Otaegui, Lisa Yeomans, and Elisabetta Boaretto. High resolution ams dates from shubayqa 1, northeast jordan reveal complex origins of late epipalaeolithic natufian in the levant. *Scientific Reports*, 7(1):17025, 2017.

[18] Amaia Arranz-Otaegui, Lara Gonzalez Carretero, Monica N. Ramsey, Dorian Q. Fuller, and Tobias Richter. Archaeobotanical evidence reveals the origins of bread 14,400 years ago in northeastern jordan. *Proceedings of the National Academy of Sciences*, 115(31):7925–7930, 2018.

[19] Li Liu, Jiajing Wang, Danny Rosenberg, Hao Zhao, Gyorgy Lengyel, and Dani Nadel. Fermented beverage and food storage in 13,000yold stone mortars at raqefet cave, israel Investigating natufian ritual feasting. *Journal of Archaeological Science: Reports*, 21:783–793, 2018.

[20] Shahal Abbo, Simcha Lev-Yadun, Manfred Heun, and Avi Gopher. On the 'lost' crops of the neolithic Near East. *Journal of Experimental Botany*, 64(4):815–822, 02 2013.

[21] A. Badr, K. M, R. Sch, H. El Rabey, S. Effgen, H. H. Ibrahim, C. Pozzi, W. Rohde, and F. Salamini. On the Origin and Domestication History of Barley (Hordeum vulgare). *Molecular Biology and Evolution*, 17(4):499–510, 04 2000.

[22] Robert J Braidwood, Jonathan D Sauer, Hans Helbaek, Paul C Mangelsdorf, Hugh C Cutler, Carleton S Coon, Ralph Linton, Julian Steward, and A Leo Oppenheim. Symposium: Did Man Once Live by Beer Alone? *American Anthropologist*, 55(4):515–526, 1953.

[23] Peter Damerow. Sumerian beer: the origins of brewing technology in ancient Mesopotamia. *Cuneiform Digital Library Journal*, 2012.

[24] Neil MacGregor. *A History of the World in 100 Objects*. 2010.

[25] J. J. Mark. The Hymn to Ninkasi, Goddess of Beer, 2011.

[26] Charlotte Beck. The Oxford companion to archaeology. *Choice Reviews Online*, 2013.

[27] Nadine Guilhou. *Myth of the Heavenly Cow*. 2010.

[28] J. J. Mark. Beer in ancient egypt, 2017.

[29] Salwa A. Maksoud, M. Nabil El Hadidi, and Wafaa Mahrous Amer. Beer from the early dynasties (3500-3400 cal B.C.) of Upper Egypt, detected by archaeochemical methods. *Vegetation History and Archaeobotany*, 2004.

[30] Max Nelson. *The barbarian's beverage: A history of beer in ancient Europe*. 2005.

[31] J. Troels-Smith, C. Jessen, and M. F. Mortensen. Modern pollen analysis and prehistoric beer - A lecture by Jørgen Troels-Smith, March 1977, 2018.

[32] Jean-Baptiste Bonnard. Male and female bodies according to Ancient Greek physicians. *Clio*, 2017.

[33] Jacques Jouanna. *Greek Medicine from Hippocrates to Galen*. 2012.

[34] M. E. Moseley, D. J. Nash, P. R. Williams, S. D. DeFrance, A. Miranda, and M. Ruales. Burning down the brewery: Establishing and evacuating an ancient imperial colony at Cerro Baul, Peru. *Proceedings of the National Academy of Sciences*, 2005.

[35] Patrick Ryan Williams, Donna J. Nash, Joshua M. Henkin, and Ruth Ann Armitage. Archaeometric Approaches to Defining Sustainable Governance: Wari Brewing Traditions and the Building of Political Relationships in Ancient Peru. *Sustainability*, 2019.

[36] Satoshi Natsume, Hiroki Takagi, Akira Shiraishi, Jun Murata, Hiromi Toyonaga, Josef Patzak, Motoshige Takagi, Hiroki Yaegashi, Aiko Uemura, Chikako Mitsuoka, Kentaro Yoshida, Karel Krofta, Honoo Satake, Ryohei Terauchi, and Eiichiro Ono. The draft genome of hop (humulus lupulus), an essence for brewing. *Plant and Cell Physiology*, 2015.

[37] Michael Moir. Hops — A Millennium Review. *Journal of American Society of Brewing Chemists*, 2000.

[38] Brian Gibson and Gianni Liti. Saccharomyces pastorianus: Genomic insights inspiring innovation for industry. *Yeast*, 2015.

[39] Mika Laitinen. Viking age brew : the craft of brewing sahti farmhouse ale, 2019.

[40] Clifford Dobell. A protozoological bicentenary: Antony van Leeuwenhoek (1632-1723) and Louis Joblot (1645-1723). *Parasitology*, 15(3):308–319, 1923.

[41] Nick Lane. The unseen World: Reflections on Leeuwenhoek (1677) 'Concerning little animals', 2015.

[42] Kutluay Uluç, Gregory C. Kujoth, and Mustafa K. Başkaya. Operating microscopes: past, present, and future. *Neurosurgical Focus*, page E4, 2009.

[43] Robert Hooke. *Micrographia or, some physiological descriptions of minute bodies made by magnifying glasses.* 1665.

[44] Lesley A. Robertson. van Leeuwenhoek microscopes-Where are they now? *FEMS Microbiology Letters*, 362(9), 2015.

[45] Brian J. Ford. Living Images from the Birth of Microscopy. *Microscopy Today*, 22(04):18–23, 2014.

[46] Antoni van Leeuwenhoek. Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infus. *Philosophical Transactions of the Royal Society of London*, 1677.

[47] A. Chaston Chapman. THE YEAST CELL: WHAT DID LEEUWENHOECK SEE ? *Journal of the Institute of Brewing*, 1931.

[48] Charles Cagniard-Latour. Mémoire sur la fermentation alcoolique. *Annales de chimie et de physique*, 68:206–222, 1838.

[49] Theodor Schwann. Vorlaufige Mittheilung betreffend Versuche uber die Weingahrung und Faulniss. *Ann. Phys. Chem.*, 41:184–93, 1837.

[50] Louis Rosenfeld. Justus Liebig and animal chemistry, 2003.

[51] Kendall A. Smith. Louis Pasteur, the father of immunology?, 2012.

[52] Louis Pasteur. Memoire sur les corpuscule organises qui existent dans latmosphere. examen de la doctrine des generations spontanese. *Annales des Sciences Naturelles*, 16:5–68, 1861.

[53] Louis Pasteur. Mémoire sur la fermentation alcoolique. *Annales de chimie et de physique*, 58:1–106, 1860.

[54] Anita Krisko and Miroslav Radman. Biology of extreme radiation resistance: The way of Deinococcus radiodurans. *Cold Spring Harbor Perspectives in Biology*, 2013.

[55] Patrick Chain, Jane Lamerdin, Frank Larimer, Warren Regala, Victoria Lao, Miriam Land, Loren Hauser, Alan Hooper, Martin Klotz, Jeanette Norton, Luis Sayavedra-Soto, Dave Arciero, Norman Hommes, Mark Whittaker, and Daniel Arp. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea. *Journal of Bacteriology*, 2003.

[56] Jose L. Adrio and Arnold L. Demain. Microbial enzymes: tools for biotechnological processes, 2014.

[57] Ljubica Vojcic, Christian Pitzler, Georgette Körfer, Felix Jakob, Ronny Martinez, Karl Heinz Maurer, and Ulrich Schwaneberg. Advances in protease engineering for laundry detergents, 2015.

[58] Fanqiang Meng, Rui Chen, Xiaoyu Zhu, Yingjian Lu, Ting Nie, Fengxia Lu, and Zhaoxin Lu. Newly Effective Milk-Clotting Enzyme from Bacillus subtilis and Its Application in Cheese Making. *Journal of Agricultural and Food Chemistry*, 2018.

[59] Michael E. Hibbing, Clay Fuqua, Matthew R. Parsek, and S. Brook Peterson. Bacterial competition: Surviving and thriving in the microbial jungle, 2010.

[60] Alexander Fleming. On the antibacterial action of cultures of a Penicillium, with special reference to their use in the isolation of B. injluenzae. *British Journal of Experimental Pathology,*, 1929.

[61] C. M. Visagie, J. Houbraken, J. C. Frisvad, S. B. Hong, C. H.W. Klaassen, G. Perrone, K. A. Seifert, J. Varga, T. Yaguchi, and R. A. Samson. Identification and nomenclature of the genus Penicillium. *Studies in Mycology*, 2014.

[62] Rustam I. Aminov. A brief history of the antibiotic era: Lessons learned and challenges for the future. *Frontiers in Microbiology*, 2010.

[63] Kate Gould. Antibiotics: from prehistory to the present day. *Journal of Antimicrobial Chemotherapy*, 71(3):572–575, 02 2016.

[64] Roswell Quinn. Rethinking antibiotic research and development: World War II and the penicillin collaborative. *American Journal of Public Health*, 2013.

[65] Gilbert Shama and Jonathan Reinarz. Allied intelligence reports on wartime German penicillin research and production. *Historical Studies in the Physical and Biological Sciences*, 2004.

[66] John Durbin Husher. *Life and Death The History of Overcoming Disease and What It Tells Us About Our Present Increasing Life Expectancy As a Result of Present Day Actions.* iUniverse Inc., 2015.

[67] American Chemical Society. Discovery and Development of Penicillin, 1999.

[68] Robert Gaynes. The discovery of penicillin—new insights after more than 75 years of clinical use. *Emerging Infectious Diseases*, 2017.

[69] Review on Antimicrobial Resistance. *Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations.* Review on Antimicrobial Resistance, 2014.

[70] A. Goffeau, G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 1996.

[71] James C. Liao, Luo Mi, Sammy Pontrelli, and Shanshan Luo. Fuelling the future: Microbial engineering for the production of sustainable biofuels, 2016.

[72] Stephanie Galanie, Kate Thodey, Isis J. Trenchard, Maria Filsinger Interrante, and Christina D. Smolke. Complete biosynthesis of opioids in yeast. *Science*, 2015.

[73] Jackson Peter, Matteo De Chiara, Anne Friedrich, Jia-Xing Yue, David Pflieger, Anders Bergström, Anastasie Sigwalt, Benjamin Barre, Kelle Freel, Agnès Llored, Corinne Cruaud, Karine Labadie, Jean-Marc Aury, Benjamin Istace, Kevin Lebrigand, Pascal Barbry, Stefan Engelen, Arnaud Lemainque, Patrick Wincker, Gianni Liti, and Joseph Schacherer. Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature*, 556(7701):339–344, 2018.

[74] James E DiCarlo, Julie E Norville, Prashant Mali, Xavier Rios, John Aach, and George M Church. Genome engineering in saccharomyces cerevisiae using crispr-cas systems. *Nucleic acids research*, 41(7):4336–4343, 04 2013.

[75] Matthew Cobb. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology*, 2017.

[76] Charles E Cook, Mary Todd Bergman, Robert D Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The european bioinformatics institute in 2016: Data growth and integration. *Nucleic acids research*, 44(D1):D20–D26, 01 2016.

[77] Aaron Golden, S. George Djorgovski, and John M. Greally. Astrogenomics: big data, old problems, old solutions? *Genome Biology*, 14(8):129, 2013.

[78] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage ϕx174 dna. *Nature*, 265(5596):687–695, 1977.

[79] Eric Lander, Cong Chen, Lauren Linton, Bruce Birren, Chad Nusbaum, Michael Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gaige, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, and Lee Rowen. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 02 2001.

[80] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu,

Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[81] Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. DNA sequencing at 40: Past, present and future. *Nature*, 2017.

[82] Felix Quitterer, Anja List, Wolfgang Eisenreich, Adelbert Bacher, and Michael Groll. Crystal structure of methylornithine synthase (pylb): Insights into the pyrrolysine biosynthesis. *Angewandte Chemie International Edition*, 51(6):1339–1342, 2012.

[83] Marco Mariotti, Gustavo Salinas, Toni Gabaldón, and Vadim N. Gladyshev. Utilization of selenocysteine in early-branching fungal phyla. *Nature Microbiology*, 4(5):759–765, 2019.

[84] Vyacheslav M Labunskyy, Dolph L Hatfield, and Vadim N Gladyshev. Selenoproteins: molecular pathways and physiological roles. *Physiological reviews*, 94(3):739–777, 07 2014.

[85] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[86] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[87] Osamu Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705 – 708, 1982.

[88] Stephen F. Altschul and Bruce W. Erickson. Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, 48(5):603 – 616, 1986.

[89] Osamu Gotoh. Optimal sequence alignment allowing for long gaps. *Bulletin of Mathematical Biology*, 52(3):359 – 373, 1990.

[90] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595, 03 2010.

[91] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963–e112963, 11 2014.

[92] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1):238, 2012.

[93] Heng Li. "minimap2: pairwise alignment for nucleotide sequences". *Bioinformatics*, 34(18):3094–3100, may 2018.

[94] Josip Marić, Ivan Sović, Krešimir Križanović, Niranjan Nagarajan, and Mile Šikić. Graphmap2 - splice-aware RNA-seq mapper for long reads. *bioRxiv*, page 720458, jan 2019.

[95] Michael S Waterman. Efficient sequence alignment algorithms. *Journal of Theoretical Biology*, 108(3):333–337, 1984.

[96] Webb Miller and Eugene W Myers. Sequence comparison with concave weighting functions. *Bulletin of Mathematical Biology*, 50(2):97–120, 1988.

[97] Reed A Cartwright. Logarithmic gap costs decrease alignment accuracy. *BMC Bioinformatics*, 7(1):527, 2006.

[98] D J Lipman and W R Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435 LP – 1441, mar 1985.

[99] W J Wilbur and D J Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the United States of America*, 80(3):726–730, 02 1983.

[100] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[101] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[102] Jodi A. Lindsay and Matthew T. G. Holden. Staphylococcus aureus: superbug, super genome? *Trends in Microbiology*, 12(8):378–385, 2020/02/16 2004.

[103] Joel Armstrong, Ian T Fiddes, Mark Diekhans, and Benedict Paten. Whole-genome alignment and comparative annotation. *Annual review of animal biosciences*, 7:41–64, 02 2019.

[104] Arthur L. Delcher, Simon Kasif, Robert D. Fleischmann, Jeremy Peterson, Owen White, and Steven L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 01 1999.

[105] David Sankoff and Peter H. Sellers. Shortcuts, diversions, and maximal chainsin partially ordered sets. *Discrete Mathematics*, 4(3):287 – 293, 1973.

[106] Michael Roberts, Wayne Hayes, Brian R. Hunt, Stephen M. Mount, and James A. Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369, 07 2004.

[107] Will P. M. Rowe. When the levee breaks: a practical guide to sketching algorithms for processing the flood of genomic data. *Genome Biology*, 20(1):199, 2019.

[108] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 09 2010.

[109] Alice Maria Giani, Guido Roberto Gallo, Luca Gianfranceschi, and Giulio Formenti. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 2019.

[110] Steven N Evans, Valerie Hower, and Lior Pachter. Coverage statistics for sequence census methods. *BMC Bioinformatics*, 11(1):430, 2010.

[111] Eric S Lander and Michael S Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239, 1988.

[112] F Sanger, A R Coulson, G F Hong, D F Hill, and G B Petersen. Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, 162(4):729–773, 1982.

[113] T R Gingeras, J P Milazzo, D Sciaky, and R J Roberts. Computer programs for the assembly of DNA sequences. *Nucleic Acids Research*, 7(2):529–543, sep 1979.

[114] R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, jun 1979.

[115] R Staden. A mew computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Research*, 8(16):3673–3694, aug 1980.

[116] Eugene W Myers. A history of DNA sequence assembly. *it - Information Technology*, 58, 2016.

[117] Hannu Peltola, Hans Söderlund, and Esko Ukkonen. SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research*, 12(1Part1):307–321, jan 1984.

[118] John Kececioglu and Eugene W Myers. Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13:7–51, feb 1995.

[119] Eugene W Myers. Toward Simplifying and Accurately Formulating Fragment Assembly. *Journal of Computational Biology*, 2(2):275–290, jan 1995.

[120] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21(suppl_2):ii79–ii85, sep 2005.

[121] Yu Lin, Jeffrey Yuan, Mikhail Kolmogorov, Max W Shen, Mark Chaisson, and Pavel A Pevzner. Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences*, 113(52):E8396—-E8405, 2016.

[122] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, 2013.

[123] Jared C Roach, Cecilie Boysen, Kai Wang, and Leroy Hood. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 26(2):345–353, 1995.

[124] James L Weber and Eugene W Myers. Human whole-genome shotgun sequencing. *Genome Research*, 7(5):401–409, May 1997.

[125] Eugene W. Myers, Granger G. Sutton, Art L. Delcher, Ian M. Dew, Dan P. Fasulo, Michael J. Flanigan, Saul A. Kravitz, Clark M. Mobarry, Knut H.J. Reinert, Karin A. Remington, Eric L. Anson, Randall A. Bolanos, Hui Hsien Chou, Catherine M. Jordan, Aaron L. Halpern, Stefano Lonardi, Ellen M. Beasley, Rhonda C. Brandon, Lin Chen, Patrick J. Dunn, Zhongwu Lai, Yong Liang, Deborah R. Nusskern, Ming Zhan, Qing Zhang, Xiangqun Zheng, Gerald M. Rubin, Mark D. Adams, and J. Craig Venter. A whole-genome assembly of Drosophila, 2000.

[126] Eugene W Myers. Toward Simplifying and Accurately Formulating Fragment Assembly. *Journal of Computational Biology*, 2(2):275–290, jan 1995.

[127] Pavel A Pevzner. 1-Tuple DNA Sequencing: Computer Analysis. *Journal of Biomolecular Structure and Dynamics*, 7(1):63–73, aug 1989.

[128] Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications*, 1:1000106, 2014.

[129] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. Allpaths: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–820, 05 2008.

[130] Daniel R Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, may 2008.

[131] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991, 2011.

[132] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748 LP – 9753, aug 2001.

[133] Dana C Crawford and Deborah A Nickerson. Definition and Clinical Importance of Haplotypes. *Annual Review of Medicine*, 56(1):303–320, sep 2004.

[134] Ehsan Motazedi, Chris Maliepaard, Richard Finkers, Richard Visser, and Dick de Ridder. Family-Based Haplotype Estimation and Allele Dosage Correction for Polyploids Using Short Sequence Reads. *Frontiers in Genetics*, 10:335, 2019.

[135] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy P L Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 36(12):1174–1182, 2018.

[136] Jan Steensels, Tim Snoek, Esther Meersman, Martina Picca Nicolino, Karin Voordeckers, and Kevin J Verstrepen. Improving industrial yeast strains: exploiting natural and artificial diversity. *FEMS microbiology reviews*, 38(5):947–995, 09 2014.

[137] M van den Broek, I Bolat, J F Nijkamp, E Ramos, M A H Luttik, F Koopman, J M Geertman, D de Ridder, J T Pronk, and J-M Daran. Chromosomal copy number variation in saccharomyces pastorianus is evidence for extensive genome dynamics in industrial lager brewing strains. *Applied and environmental microbiology*, 81(18):6253–6267, 09 2015.

[138] Ehsan Motazedi, Dick de Ridder, Richard Finkers, Samantha Baldwin, Susan Thomson, Katrina Monaghan, and Chris Maliepaard. TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics*, 34(22):3864–3872, jun 2018.

[139] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226–232, 01 2012.

[140] Rei Kajitani, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, Masayuki Harada, Eiji Nagayasu, Haruhiko Maruyama, Yuji Kohara, Asao Fujiyama, Tetsuya Hayashi, and Takehiko Itoh. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, 24(8):1384–1395, 08 2014.

[141] Jurgen F. Nijkamp, Marcel van den Broek, Erwin Datema, Stefan de Kok, Lizanne Bosman, Marijke A. Luttik, Pascale Daran-Lapujade, Wanwipa Vongsangnak, Jens Nielsen, Wilbert HM Heijne, Paul Klaassen, Chris J. Paddon, Darren Platt, Peter Kötter, Roeland C. van Ham, Marcel JT Reinders, Jack T. Pronk, Dick de Ridder, and Jean-Marc Daran. De novo sequencing, assembly and analysis of the genome of the laboratory strain saccharomyces cerevisiae cen.pk113-7d, a model for modern industrial biotechnology. *Microbial Cell Factories*, 11(1):36, 2012.

[142] Pooja K Strope, Daniel A Skelly, Stanislav G Kozmin, Gayathri Mahadevan, Eric A Stone, Paul M Magwene, Fred S Dietrich, and John H McCusker. The 100-genomes strains, an s. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome research*, 25(5):762–774, 05 2015.

[143] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt, and Ali Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, 12(8):780–786, 2015.

[144] Jeong-Sun Seo, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, Han Cao, Ji-Young Yun, Jihye Kim, Junho Kuk, Gun Hwa Park, Juhyeok Kim, Hanna Ryu, Jongbum Kim, Mira Roh, Jeonghun Baek, Michael W Hunkapiller, Jonas Korlach, Jong-Yeon Shin, and Changhoon Kim. De novo assembly and phasing of a Korean human genome. *Nature*, 538(7624):243–247, 2016.

[145] Sergey Koren and Adam M Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23:110–120, 2015.

[146] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[147] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, 2018.

[148] Jason D Merker, Aaron M Wenger, Tam Sneddon, Megan Grove, Zachary Zappala, Laure Fresard, Daryl Waggott, Sowmi Utiramerur, Yanli Hou, Kevin S Smith, Stephen B Montgomery, Matthew Wheeler, Jillian G Buchan, Christine C Lambert, Kevin S Eng, Luke Hickey, Jonas Korlach, James Ford, and Euan A Ashley. Long-read genome sequencing identifies causal structural variation in a mendelian disease. *Genetics in Medicine*, 20(1):159–163, 2018.

[149] Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, 2018.

[150] Rachael E. Workman, Alison D. Tang, Paul S. Tang, Miten Jain, John R. Tyson, Roham Razaghi, Philip C. Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen L. Jones, Cameron M. Soulette, Terrance P. Snutch, Nicholas Loman, Benedict Paten, Matthew Loose, Jared T. Simpson, Hugh E. Olsen, Angela N. Brooks, Mark Akeson, and Winston Timp. Nanopore native rna sequencing of a human poly(a) transcriptome. *Nature Methods*, 16(12):1297–1305, 2019.

[151] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, may 2017.

[152] Chen-Shan Chin and Asif Khalak. Human Genome Assembly in 100 Minutes. *bioRxiv*, page 705616, jan 2019.

[153] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, 2015.

[154] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[155] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics (Oxford, England)*, 32(14):2103–2110, 07 2016.

[156] Ivan Sović, Mile Šikić, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen, and Niranjan Nagarajan. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications*, 7(1):11307, 2016.

[157] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):2542, 2018.

[158] Gene Myers. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, 46(3):395–415, May 1999.

[159] Michael Farrar. Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics*, 23(2):156–161, 11 2006.

[160] K. Benkrid, Y. Liu, and A. Benkrid. A highly parameterized and efficient fpga-based skeleton for pairwise biological sequence alignment. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 17(4):561–570, April 2009.

[161] Sorin Istrail, Granger G. Sutton, Liliana Florea, Aaron L. Halpern, Clark M. Mobarry, Ross Lippert, Brian Walenz, Hagit Shatkay, Ian Dew, Jason R. Miller, Michael J. Flanigan, Nathan J. Edwards, Randall Bolanos, Daniel Fasulo, Bjarni V. Halldorsson, Sridhar Hannenhalli, Russell Turner, Shibu Yooseph, Fu Lu, Deborah R. Nusskern, Bixiong Chris Shue, Xiangqun Holly Zheng, Fei Zhong, Arthur L. Delcher, Daniel H. Huson, Saul A. Kravitz, Laurent Mouchard, Knut Reinert, Karin A. Remington, Andrew G. Clark, Michael S. Waterman, Evan E. Eichler, Mark D. Adams, Michael W. Hunkapiller, Eugene W. Myers, and J. Craig Venter. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proceedings of the National Academy of Sciences*, 101(7):1916–1921, 2004.

[162] Dmitry Antipov, Anton Korobeynikov, Jeffrey S. McLean, and Pavel A. Pevzner. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015, 11 2015.

[163] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29, June 1997.

[164] Ondrej Chum, James Philbin, and Andrew Zisserman. *Near Duplicate Image Detection: min-Hash and tf-idf Weighting.* jan 2008.

[165] Jason R Miller, Arthur L Delcher, Sergey Koren, Eli Venter, Brian P Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824, oct 2008.

[166] Kishwar Shafin, Trevor Pesout, Ryan Lorig-Roach, Marina Haukness, Hugh E Olsen, Colleen Bosworth, Joel Armstrong, Kristof Tigyi, Nicholas Maurer, Sergey Koren, Fritz J Sedlazeck, Tobias Marschall, Simon Mayes, Vania Costa, Justin M Zook, Kelvin J Liu, Duncan Kilburn, Melanie Sorensen, Katy M Munson, Mitchell R Vollger, Evan E Eichler, Sofie Salama, David Haussler, Richard E Green, Mark Akeson, Adam Phillippy, Karen H Miga, Paolo Carnevali, Miten Jain, and Benedict Paten. Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit. *bioRxiv*, page 715722, jan 2019.

[167] R R Wick and K E Holt. Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 1; peer review: 4 approved]. *F1000Research*, 8(2138), 2019.

[168] Robert Vaser, Ivan Sovic, Niranjan Nagarajan, and Mile Sikic. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, jan 2017.

[169] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17(2):155–158, 2020.

[170] Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R Ecker, Dario Cantu, David R Rank, and Michael C Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054, 2016.

[171] Xingtan Zhang, Ruoxi Wu, Yibin Wang, Jiaxin Yu, and Haibao Tang. Unzipping haplotypes in diploid and polyploid genomes. *Computational and Structural Biotechnology Journal*, 18:66–72, 2020.

[172] Rei Kajitani, Dai Yoshimura, Miki Okuno, Yohei Minakuchi, Hiroshi Kagoshima, Asao Fujiyama, Kaoru Kubokawa, Yuji Kohara, Atsushi Toyoda, and Takehiko Itoh. Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature Communications*, 10(1):1702, 2019.

[173] Rachel M Sherman and Steven L Salzberg. Pan-genomics in the human genome era. *Nature Reviews Genetics*, 2020.

[174] Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, oct 2016.

[175] Walter M. Fitch. An improved method of testing for evolutionary homology. *Journal of Molecular Biology*, 16:9–16, 1966.

[176] Huan Fan, Anthony R. Ives, Yann Surget-Groba, and Charles H. Cannon. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, 16(1):522, 2015.

[177] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17(1):132, 2016.

[178] Phelim Bradley, Henk C. den Bakker, Eduardo P. C. Rocha, Gil McVean, and Zamin Iqbal. Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology*, 37(2):152–159, 2019.

[179] Carl R. Woese and George E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[180] F. Sanger and A.R. Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441 – 448, 1975.

[181] Pehr Edman, Erik Högfeldt, Lars Gunnar Sillén, and Per-Olof Kinell. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chemica Scandinavica*, 4:283–293, 1950.

[182] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal*, 4:463–481, 1951.

[183] F. Sanger and E. O. Thompson. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *The Biochemical journal*, 3:353–366, 1953.

[184] F. Sanger and E. O. Thompson. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *The Biochemical journal*, 3:366–374, 1953.

[185] V. M. Ingram. A specific chemical difference between the globins of normal human and sickle-cell anæmia hæmoglobin. *Nature*, 178:792–794, 1956.

[186] V. M. Ingram. Gene mutations in human hæmoglobin: The chemical difference between normal and sickle cell hæmoglobin. *Nature*, 180:326–328, 1957.

[187] J D Watson and F H C Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.

[188] F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[189] Antony O W Stretton. The first sequence: Fred Sanger and insulin. *Genetics*, 162(2):527–532, 2002.

[190] Mireia Coscolla and Sebastien Gagneux. Consequences of genomic diversity in mycobacterium tuberculosis. *Seminars in Immunology*, 26(6):431 – 444, 2014. Immunity to Mycobacterium tuberculosis.

[191] Anthony G. Tsolaki, Aaron E. Hirsh, Kathryn DeRiemer, Jose Antonio Enciso, Melissa Z. Wong, Margaret Hannan, Yves-Olivier L. Goguet de la Salmoniere, Kumiko Aman, Midori Kato-Maeda, and Peter M. Small. Functional and evolutionary genomics of mycobacterium tuberculosis: Insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences*, 101(14):4865–4870, 2004.

[192] Oksana Lukjancenko, Trudy M. Wassenaar, and David W. Ussery. Comparison of 61 sequenced escherichia coli genomes. *Microbial Ecology*, 60(4):708–720, 2010.

[193] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114, 2018.

[194] Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 07 2015.

[195] Narendrakumar M. Chaudhari, Vinod Kumar Gupta, and Chitra Dutta. Bpga- an ultra-fast pan-genome analysis pipeline. *Scientific Reports*, 6(1):24373, 2016.

[196] Chad Laing, Cody Buchanan, Eduardo N Taboada, Yongxiang Zhang, Andrew Kropinski, Andre Villegas, James E Thomas, and Victor P J Gannon. Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC bioinformatics*, 11:461–461, 09 2010.

[197] Yongbing Zhao, Xinmiao Jia, Junhui Yang, Yunchao Ling, Zhang Zhang, Jun Yu, Jiayan Wu, and Jingfa Xiao. Pangp: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics (Oxford, England)*, 30(9):1297–1299, 05 2014.

[198] Derrick E Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. Panoct: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research*, 40(22):e172–e172, 12 2012.

[199] Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418, 11 2011.

[200] Matthew N Benedict, James R Henriksen, William W Metcalf, Rachel J Whitaker, and Nathan D Price. Itep: an integrated toolkit for exploration of microbial pan-genomes. *BMC genomics*, 15:8–8, 01 2014.

[201] Ola Brynildsrud, Jon Bohlin, Lonneke Scheffer, and Vegard Eldholm. Rapid scoring of genes in microbial pan-genome-wide association studies with scoary. *Genome Biology*, 17(1):238, 2016.

[202] AndréB. Canelas, Nicola Harrison, Alessandro Fazio, Jie Zhang, Juha-Pekka Pitkänen, Joost van den Brink, Barbara M. Bakker, Lara Bogner, Jildau Bouwman, Juan I. Castrillo, Ayca Cankorur, Pramote Chumnanpuen, Pascale Daran-Lapujade, Duygu Dikicioglu, Karen van Eunen, Jennifer C. Ewald, Joseph J. Heijnen, Betul Kirdar, Ismo Mattila, Femke I. C. Mensonides, Anja Niebel, Merja Penttilä, Jack T. Pronk, Matthias Reuss, Laura Salusjärvi, Uwe Sauer, David Sherman, Martin Siemann-Herzberg, Hans Westerhoff, Johannes de Winde, Dina Petranovic, Stephen G. Oliver, Christopher T. Workman, Nicola Zamboni, and Jens Nielsen. Integrated multilaboratory systems biology reveals differences in protein metabolism between two reference yeast strains. *Nature Communications*, 1(1):145, 2010.

[203] M T A P Kresnowati, W A van Winden, M J H Almering, A ten Pierick, C Ras, T A Knijnenburg, P Daran-Lapujade, J T Pronk, J J Heijnen, and J M Daran. When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Molecular Systems Biology*, 2(1):49, 2006.

[204] J.P van Dijken, J Bauer, L Brambilla, P Duboc, J.M Francois, C Gancedo, M.L.F Giuseppin, J.J Heijnen, M Hoare, H.C Lange, E.A Madden, P Niederberger, J Nielsen, J.L Parrou, T Petit, D Porro, M Reuss, N van Riel, M Rizzi, H.Y Steensma, C.T Verrips, J Vindeløv, and J.T Pronk. An interlaboratory comparison of physiological and genetic properties of four saccharomyces cerevisiae strains. *Enzyme and Microbial Technology*, 26(9):706 – 714, 2000.

[205] Karl-Dieter Entian and Peter Kötter. 25 yeast genetic strain and plasmid collections. In Ian Stansfield and Michael JR Stark, editors, *Yeast Gene Analysis*, volume 36 of *Methods in Microbiology*, pages 629 – 666. Academic Press, 2007.

[206] JoséManuel Otero, Wanwipa Vongsangnak, Mohammad A. Asadollahi, Roberto Olivares-Hernandes, Jérôme Maury, Laurent Farinelli, Loïc Barlocher, Magne Østerås, Michel Schalk, Anthony Clark, and Jens Nielsen. Whole genome sequencing of saccharomyces cerevisiae: from genotype to phenotype for improved metabolic engineering applications. *BMC Genomics*, 11(1):723, 2010.

[207] 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, and Thomas Abeel. Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D. *FEMS Yeast Research*, 17(7), 09 2017. fox074.

[208] Miki Okuno, Rei Kajitani, Rie Ryusui, Hiroya Morimoto, Yukiko Kodama, and Takehiko Itoh. Next-generation sequencing analysis of lager brewing yeast strains reveals the evolutionary history of interspecies hybridization. *DNA Research*, 23(1):67–80, 01 2016.

[209] Arthur R. Gorter de Vries, Jack T. Pronk, and Jean-Marc G. Daran. Industrial relevance of chromosomal copy number variation in saccharomyces yeasts. *Applied and Environmental Microbiology*, 83(11), 2017.

[210] Andrea Walther, Ana Hesselbart, and Jürgen Wendland. Genome sequence of saccharomyces carlsbergensis, the world's first pure culture lager yeast. *G3 (Bethesda, Md.)*, 4(5):783–793, 02 2014.

[211] Yoshihiro Nakao, Takeshi Kanamori, Takehiko Itoh, Yukiko Kodama, Sandra Rainieri, Norihisa Nakamura, Tomoko Shimonaga, Masahira Hattori, and Toshihiko Ashikari. Genome Sequence of the Lager Brewing Yeast, an Interspecies Hybrid. *DNA Research*, 16(2):115–129, 03 2009.

[212] Alex N. Salazar, Arthur R. Gorter de Vries, Marcel van den Broek, Nick Brouwers, Pilar de la Torre Cortès, Niels G. A. Kuijpers, Jean-Marc G. Daran, and Thomas Abeel. Chromosome level assembly and comparative genome analysis confirm lager-brewing yeasts originated from a single hybridization. *BMC Genomics*, 20(1):916, 2019.

[213] Alex N. Salazar and Thomas Abeel. Alpaca: a kmer-based approach for investigating mosaic structures in microbial genomes. *bioRxiv*, 2019.

[214] Alex N. Salazar, Franklin L. Nobrega, Christine Anyansi, Cristian Aparicio-Maldonado, Ana Rita Costa, Anna C. Haagsma, Anwar Hiralal, Ahmed Mahfouz, Rebecca E. McKenzie, Teunke van Rossum, Stan J. J. Brouns, and Thomas Abeel. An educational guide for nanopore sequencing in the classroom. *PLOS Computational Biology*, 16(1):1–7, 01 2020.

[215] Elaine R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2020/02/22 2008.

[216] Pauline C Ng and Ewen F Kirkness. Whole genome sequencing. In *Genetic variation*, pages 215–226. Springer, 2010.

[217] Erwin L. van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, 2020/02/22 2014.

[218] Kinnari Matheson, Lance Parsons, and Alison Gammie. Whole-genome sequence and variant analysis of w303, a widely-used strain of saccharomyces cerevisiae. *G3: Genes, Genomes, Genetics*, 7(7):2219–2226, 2017.

[219] J. Michael Cherry, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng, and Edith D. Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705, 11 2011.

[220] Daniel González-Ramos, Arthur R. Gorter de Vries, Sietske S. Grijseels, Margo C. van Berkum, Steve Swinnen, Marcel van den Broek, Elke Nevoigt, Jean-Marc G. Daran, Jack T. Pronk, and Antonius J. A. van Maris. A new laboratory evolution

approach to select for constitutive acetic acid tolerance in saccharomyces cerevisiae and identification of causal mutations. *Biotechnology for Biofuels*, 9(1):173, 2016.

[221] Ioannis Papapetridis, Marlous van Dijk, Antonius J. A. van Maris, and Jack T. Pronk. Metabolic engineering strategies for optimizing acetate reduction, ethanol yield and osmotolerance in saccharomyces cerevisiae. *Biotechnology for Biofuels*, 10(1):107, 2017.

[222] Jurgen Nijkamp, Wynand Winterbach, Marcel van den Broek, Jean-Marc Daran, Marcel Reinders, and Dick de Ridder. Integrating genome assemblies with MAIA. *Bioinformatics*, 26(18):i433–i439, 09 2010.

[223] Fiona E Pryde, Tonie C Huckle, and Edward J Louis. 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(4):371–382, 1995.

[224] Jin M Kim, Swathi Vanguri, Jef D Boeke, Abram Gabriel, and Daniel F Voytas. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete saccharomyces cerevisiae genome sequence. *Genome research*, 8(5):464–478, 1998.

[225] Anders Bergström, Jared T. Simpson, Francisco Salinas, Benjamin Barré, Leopold Parts, Amin Zia, Alex N. Nguyen Ba, Alan M. Moses, Edward J. Louis, Ville Mustonen, Jonas Warringer, Richard Durbin, and Gianni Liti. A High-Definition View of Functional Genetic Variation from Natural Yeast Genomes. *Molecular Biology and Evolution*, 31(4):872–888, 01 2014.

[226] M Carlson, J L Celenza, and F J Eng. Evolution of the dispersed suc gene family of saccharomyces by rearrangements of chromosome telomeres. *Molecular and Cellular Biology*, 5(11):2894–2902, 1985.

[227] Gennadi I Naumov, Elena S Naumova, and Edward J Louis. Genetic mapping of the α-galactosidase mel gene family on right and left telomeres of saccharomyces cerevisiae. *Yeast*, 11(5):481–483, 1995.

[228] Paulina Jordan, Jun-Yong Choe, Eckhard Boles, and Mislav Oreb. Hxt13, hxt15, hxt16 and hxt17 from saccharomyces cerevisiae represent a novel type of polyol transporters. *Scientific Reports*, 6(1):23502, 2016.

[229] Marie-Ange Teste, Jean Marie François, and Jean-Luc Parrou. Characterization of a new multigene family encoding isomaltases in the yeast saccharomyces cerevisiae, the ima family. *Journal of Biological Chemistry*, 285(35):26815–26824, 2010.

[230] Muriel Denayrolles, Edouard Pinot de Villechenon, Aline Lonvaud-Funel, and M. Aigle. Incidence of suc-rtm telomeric repeated genes in brewing and wild wine strains of saccharomyces. *Current Genetics*, 31(6):457–461, 1997.

[231] A. W. R. H. Teunissen and H. Y. Steensma. The dominant flocculation genes of saccharomyces cerevisiae constitute a new subtelomeric gene family. *Yeast*, 11(11):1001–1013, 1995.

[232] Elizabeth J. Lodolo, Johan L.F. Kock, Barry C. Axcell, and Martin Brooks. The yeast Saccharomyces cerevisiae– the main character in beer brewing. *FEMS Yeast Research*, 8(7):1018–1036, 11 2008.

[233] Chris A. Brown, Andrew W. Murray, and Kevin J. Verstrepen. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology*, 20(10):895–903, 2020/02/22 2010.

[234] Sean J. McIlwain, David Peris, Maria Sardi, Oleg V. Moskvin, Fujie Zhan, Kevin S. Myers, Nicholas M. Riley, Alyssa Buzzell, Lucas S. Parreiras, Irene M. Ong, Robert Landick, Joshua J. Coon, Audrey P. Gasch, Trey K. Sato, and Chris Todd Hittinger. Genome sequence and analysis of a stress-tolerant, wild-derived strain of saccharomyces cerevisiae used in biofuels research. *G3: Genes, Genomes, Genetics*, 6(6):1757–1766, 2016.

[235] Francesca Giordano, Louise Aigrain, Michael A. Quail, Paul Coupland, James K. Bonfield, Robert M. Davies, German Tischler, David K. Jackson, Thomas M. Keane, Jing Li, Jia-Xing Yue, Gianni Liti, Richard Durbin, and Zemin Ning. De novo yeast genome assemblies from minion, pacbio and miseq platforms. *Scientific Reports*, 7(1):3935, 2017.

[236] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C. Schatz, and W. Richard McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25(11):1750–1756, 2015.

[237] Benjamin Istace, Anne Friedrich, Léo d'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti, Arnaud Lemainque, Stefan Engelen, Patrick Wincker, Joseph Schacherer, and Jean-Marc Aury. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience*, 6(2), 01 2017. giw018.

[238] 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, and Herman P Spaink. De novo whole-genome assembly of a wild type yeast isolate using nanopore sequencing. *F1000Research*, 6:618–618, 05 2017.

[239] Nicholas J. Loman and Aaron R. Quinlan. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, 30(23):3399–3401, 08 2014.

[240] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 06 2009.

[241] Stefan Kurtz, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12, 2004.

[242] Carson Holt and Mark Yandell. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1):491, 2011.

[243] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[244] Iain Milne, Gordon Stephen, Micha Bayer, Peter J.A. Cock, Leighton Pritchard, Linda Cardle, Paul D. Shaw, and David Marshall. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, 14(2):193–202, 03 2012.

[245] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[246] Pascale Daran-Lapujade, Jean-Marc Daran, Peter Kötter, Thomas Petit, Matthew DW Piper, and Jack T Pronk. Comparative genotyping of the saccharomyces cerevisiae laboratory strains s288c and cen. pk113-7d using oligonucleotide microarrays. *FEMS yeast research*, 4(3):259–269, 2003.

[247] Jaap Venema and David Tollervey. Ribosome synthesis in saccharomyces cerevisiae. *Annual Review of Genetics*, 33(1):261–311, 1999. PMID: 10690410.

[248] Kenneth H. Wolfe and Denis C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997.

[249] Pascale Daran-Lapujade, Jean-Marc Daran, Marijke AH Luttik, Marinka JH Almering, Jack T Pronk, and Peter Kötter. An atypical pmr2 locus is responsible for hypersensitivity to sodium and lithium cations in the laboratory strain saccharomyces cerevisiae cen. pk113-7d. *FEMS yeast research*, 9(5):789–792, 2009.

[250] G. Fischer, S. A. James, I. N. Roberts, S. G. Oliver, and E. J. Louis. Chromosomal evolution in saccharomyces. *Nature*, 405(6785):451–454, 2000.

[251] Jasmine M. Bracher, Erik de Hulster, Charlotte C. Koster, Marcel van den Broek, Jean-Marc G. Daran, Antonius J. A. van Maris, and Jack T. Pronk. Laboratory evolution of a biotin-requiring saccharomyces cerevisiae strain for full biotin prototrophy and identification of causal mutations. *Applied and Environmental Microbiology*, 83(16), 2017.

[252] Matthew Loose, Sunir Malla, and Michael Stout. Real-time selective sequencing using nanopore technology. *Nature Methods*, 13(9):751–754, 2016.

[253] Samina Naseeb, Zorana Carter, David Minnis, Ian Donaldson, Leo Zeef, and Daniela Delneri. Widespread Impact of Chromosomal Inversions on Gene Expression Uncovers Robustness via Phenotypic Buffering. *Molecular Biology and Evolution*, 33(7):1679–1696, 02 2016.

[254] Bart Oud, Antonius J. A. van Maris, Jean-Marc Daran, and Jack T. Pronk. Genome-wide analytical approaches for reverse metabolic engineering of industrially relevant phenotypes in yeast. *FEMS Yeast Research*, 12(2):183–196, 03 2012.

[255] Howard Ochman, Jeffrey G. Lawrence, and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

[256] Sandra Rainieri, Yukiko Kodama, Yoshinobu Kaneko, Kozaburo Mikata, Yoshihiro Nakao, and Toshihiko Ashikari. Pure and mixed genetic lines of saccharomyces bayanus and saccharomyces pastorianus and their contribution to the lager brewing strain genome. *Applied and environmental microbiology*, 72(6):3968–3974, 06 2006.

[257] William Martin. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays*, 21(2):99–104, 1999.

[258] Bernard A Dujon and Edward J Louis. Genome diversity and evolution in the budding yeasts (saccharomycotina). *Genetics*, 206(2):717–750, 06 2017.

[259] Juan Germán Rodríguez, Camilo Pino, Andreas Tauch, and Martha Isabel Murcia. Complete genome sequence of the clinical beijing-like strain mycobacterium tuberculosis 323 using the pacbio real-time sequencing platform. *Genome announcements*, 3(2):e00371–15, 04 2015.

[260] Jia-Xing Yue, Jing Li, Louise Aigrain, Johan Hallin, Karl Persson, Karen Oliver, Anders Bergström, Paul Coupland, Jonas Warringer, Marco Cosentino Lagomarsino, Gilles Fischer, Richard Durbin, and Gianni Liti. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics*, 49(6):913–924, 2017.

[261] Abigail L Manson, Keira A Cohen, Thomas Abeel, Christopher A Desjardins, Derek T Armstrong, 3rd Barry, Clifton E, Jeannette Brand, TBResist Global Genome Consortium, Sinéad B Chapman, Sang-Nae Cho, Andrei Gabrielian, James Gomez, Andreea M Jodals, Moses Joloba, Pontus Jureen, Jong Seok Lee, Lesibana Malinga, Mamoudou Maiga, Dale Nordenberg, Ecaterina Noroc, Elena Romancenco, Alex Salazar, Willy Ssengooba, A A Velayati, Kathryn Winglee, Aksana Zalutskaya, Laura E Via, Gail H Cassell, Susan E Dorman, Jerrold Ellner, Parissa Farnia, James E Galagan, Alex Rosenthal, Valeriu Crudu, Daniela Homorodean, Po-Ren Hsueh, Sujatha Narayanan, Alexander S Pym, Alena Skrahina, Soumya Swaminathan, Martie Van der Walt, David Alland, William R Bishai, Ted Cohen, Sven Hoffner, Bruce W Birren, and Ashlee M Earl. Genomic analysis of globally diverse mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nature genetics*, 49(3):395–402, 03 2017.

[262] Diego Libkind, Chris Todd Hittinger, Elisabete Valério, Carla Gonçalves, Jim Dover, Mark Johnston, Paula Gonçalves, and José Paulo Sampaio. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences*, 108(35):14539–14544, 2011.

[263] Kristoffer Krogerus, Richard Preiss, and Brian Gibson. A unique saccharomyces cerevisiae ×saccharomyces uvarum hybrid isolated from norwegian farmhouse beer: Characterization and reconstruction. *Frontiers in microbiology*, 9:2253–2253, 09 2018.

[264] Pedro Almeida, Carla Gonçalves, Sara Teixeira, Diego Libkind, Martin Bontrager, Isabelle Masneuf-Pomarède, Warren Albertin, Pascal Durrens, David James Sherman, Philippe Marullo, Chris Todd Hittinger, Paula Gonçalves, and JoséPaulo Sampaio. A gondwanan imprint on global diversity and domestication of wine and cider yeast saccharomyces uvarum. *Nature Communications*, 5(1):4044, 2014.

[265] David Peris, Quinn K. Langdon, Ryan V. Moriarty, Kayla Sylvester, Martin Bontrager, Guillaume Charron, Jean-Baptiste Leducq, Christian R. Landry, Diego Libkind, and Chris Todd Hittinger. Complex ancestries of lager-brewing hybrids were shaped by standing variation in the wild yeast saccharomyces eubayanus. *PLOS Genetics*, 12(7):1–20, 07 2016.

[266] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. Structural variant calling: the long and the short of it. *Genome Biology*, 20(1):246, 2019.

[267] Yukiko Kodama, Morten C. Kielland-Brandt, and Jørgen Hansen. *Lager brewing yeast*, pages 145–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[268] S. Dequin. The potential of genetic engineering for improving brewing, wine-making and baking yeasts. *Applied Microbiology and Biotechnology*, 56(5):577–588, 2001.

[269] Barbara Dunn and Gavin Sherlock. Reconstruction of the genome origins and evolution of the hybrid lager yeast saccharomyces pastorianus. *Genome Research*, 18(10):1610–1623, 2008.

[270] Miguel de Barros Lopes, Jennifer R. Bellon, Neil J. Shirley, and Philip F. Ganter. Evidence for multiple interspecific hybridization in Saccharomyces sensu stricto species. *FEMS Yeast Research*, 1(4):323–331, 01 2002.

[271] Marit Hebly, Anja Brickwedde, Irina Bolat, Maureen R.M. Driessen, Erik A.F. de Hulster, Marcel van den Broek, Jack T. Pronk, Jan-Maarten Geertman, Jean-Marc Daran, and Pascale Daran-Lapujade. S. cerevisiae × S. eubayanus interspecific hybrid, the best of both worlds and beyond. *FEMS Yeast Research*, 15(3), 03 2015. fov005.

[272] Kristoffer Krogerus, Frederico Magalhães, Virve Vidgren, and Brian Gibson. New lager yeast strains generated by interspecific hybridization. *Journal of Industrial Microbiology & Biotechnology*, 42(5):769–778, 2015.

[273] Sarah K Hewitt, Ian J Donaldson, Simon C Lovell, and Daniela Delneri. Sequencing and characterisation of rearrangements in three s. pastorianus strains reveals the presence of chimeric genes and gives evidence of breakpoint reuse. *PloS one*, 9(3):e92203–e92203, 03 2014.

[274] Gianni Liti, Antonella Peruffo, Steve A. James, Ian N. Roberts, and Edward J. Louis. Inferences of evolutionary relationships from a population survey of ltr-retrotransposons and telomeric-associated sequences in the saccharomyces sensu stricto complex. *Yeast*, 22(3):177–192, 2005.

[275] Chandre Monerawela, Tharappel C. James, Kenneth H. Wolfe, and Ursula Bond. Loss of lager specific genes and subtelomeric regions define two different Saccharomyces cerevisiae lineages for Saccharomyces pastorianus Group I and II strains. *FEMS Yeast Research*, 15(2), 02 2015. fou008.

[276] Kristoffer Krogerus, Mikko Arvas, Matteo De Chiara, Frederico Magalhães, Laura Mattinen, Merja Oja, Virve Vidgren, Jia-Xing Yue, Gianni Liti, and Brian Gibson. Ploidy influences the functional attributes of de novo lager yeast hybrids. *Applied Microbiology and Biotechnology*, 100(16):7203–7222, 2016.

[277] EmilyClare Baker, Bing Wang, Nicolas Bellora, David Peris, Amanda Beth Hulfachor, Justin A. Koshalek, Marie Adams, Diego Libkind, and Chris Todd Hittinger. The Genome Sequence of Saccharomyces eubayanus and the Domestication of Lager-Brewing Yeasts. *Molecular Biology and Evolution*, 32(11):2818–2831, 08 2015.

[278] Arthur R. Gorter de Vries, Maaike A. Voskamp, Aafke C. A. van Aalst, Line H. Kristensen, Liset Jansen, Marcel van den Broek, Alex N. Salazar, Nick Brouwers, Thomas Abeel, Jack T. Pronk, and Jean-Marc G. Daran. Laboratory evolution of a saccharomyces cerevisiae × s. eubayanus hybrid under simulated lager-brewing conditions. *Frontiers in Genetics*, 10:242, 2019.

[279] Anja Brickwedde, Nick Brouwers, Marcel van den Broek, Joan S. Gallego Murillo, Julie L. Fraiture, Jack T. Pronk, and Jean-Marc G. Daran. Structural, physiological and regulatory analysis of maltose transporter genes in saccharomyces eubayanus cbs 12357t. *Frontiers in Microbiology*, 9:1786, 2018.

[280] Jian Bing, Pei-Jie Han, Wan-Qiu Liu, Qi-Ming Wang, and Feng-Yan Bai. Evidence for a far east asian origin of lager beer yeast. *Current Biology*, 24(10):R380–R381, 2020/02/22 2014.

[281] Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 01 1999.

[282] Sebastiaan E. Van Mulders, Maarten Ghequire, Luk Daenen, Pieter J. Verbelen, Kevin J. Verstrepen, and Freddy R. Delvaux. Flocculation gene variability in industrial brewer's yeast strains. *Applied Microbiology and Biotechnology*, 88(6):1321–1331, 2010.

[283] Osamu Kobayashi, Nobuyuki Hayashi, Ryota Kuroki, and Hidetaka Sone. Region of flo1 proteins responsible for sugar recognition. *Journal of bacteriology*, 180(24):6503–6510, 1998.

[284] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 04 2014.

[285] Sergio L. Alves, Ricardo A. Herberts, Claudia Hollatz, Debora Trichez, Luiz C. Miletti, Pedro S. de Araujo, and Boris U. Stambuk. Molecular analysis of maltotriose active transport and fermentation by saccharomyces cerevisiae reveals a determinant role for the agt1 permease. *Applied and Environmental Microbiology*, 74(5):1494–1501, 2008.

[286] Y S Chang, R A Dubin, E Perkins, C A Michels, and R B Needleman. Identification and characterization of the maltose permease in genetically defined saccharomyces strain. *Journal of Bacteriology*, 171(11):6148–6154, 1989.

[287] Gennadi I Naumov, Elena S Naumova, and CA Michels. Genetic variation of the repeated mal loci in natural populations of saccharomyces cerevisiae and saccharomyces paradoxus. *Genetics*, 136(3):803–812, 1994.

[288] C R Zastrow, C. Hollatz, P S de Araujo, and B U Stambuk. Maltotriose fermentation by saccharomyces cerevisiae. *Journal of Industrial Microbiology and Biotechnology*, 27(1):34–38, 2001.

[289] EmilyClare P. Baker and Chris Todd Hittinger. Evolution of a novel chimeric maltotriose transporter in saccharomyces eubayanus from parent proteins unable to perform this function. *PLOS Genetics*, 15(4):1–23, 04 2019.

[290] Sebastiaan E. Van Mulders, Els Christianen, Sofie M.G. Saerens, Luk Daenen, Pieter J. Verbelen, Ronnie Willaert, Kevin J. Verstrepen, and Freddy R. Delvaux. Phenotypic diversity of Flo protein family-mediated adhesion in Saccharomyces cerevisiae. *FEMS Yeast Research*, 9(2):178–190, 02 2009.

[291] BL Miki, N HUNG Poon, ALLAN P James, and VERNER L Seligy. Possible mechanism for flocculation interactions governed by gene flo1 in saccharomyces cerevisiae. *Journal of Bacteriology*, 150(2):878–889, 1982.

[292] Pascale B Dengis, LR Nelissen, and Paul G Rouxhet. Mechanisms of yeast flocculation: comparison of top-and bottom-fermenting strains. *Appl. Environ. Microbiol.*, 61(2):718–728, 1995.

[293] Manuel Fidalgo, Ramon R. Barrales, and Juan Jimenez. Coding repeat instability in the flo11 gene of saccharomyces yeasts. *Yeast*, 25(12):879–889, 2008.

[294] Giacomo Zara, Severino Zara, Claudia Pinna, Salvatore Marceddu, and Marilena Budroni. Flo11 gene length and transcriptional level affect biofilm-forming ability of wild flor strains of saccharomyces cerevisiae. *Microbiology*, 155(12):3838–3846, 2009.

[295] Kevin J Verstrepen, An Jansen, Fran Lewitter, and Gerald R Fink. Intragenic tandem repeats generate functional variability. *Nature Genetics*, 37(9):986–990, 2005.

[296] Nan Liu, Dongli Wang, Zhao Yue Wang, Xiu Ping He, and Borun Zhang. Genetic basis of flocculation phenotype conversion in Saccharomyces cerevisiae. *FEMS Yeast Research*, 7(8):1362–1370, 12 2007.

[297] T. Ogata, M. Izumikawa, K. Kohno, and K. Shibata. Chromosomal location of lg-flo1 in bottom-fermenting yeast and the flo5 locus of industrial yeast. *Journal of Applied Microbiology*, 105(4):1186–1198, 2008.

[298] E.V. Soares. Flocculation in saccharomyces cerevisiae: a review. *Journal of Applied Microbiology*, 110(1):1–18, 2011.

[299] Ping Li, Xuewu Guo, Tingting Shi, Zhihui Hu, Yefu Chen, Liping Du, and Dongguang Xiao. Reducing diacetyl production of wine by overexpressing bdh1 and bdh2 in saccharomyces uvarum. *Journal of Industrial Microbiology & Biotechnology*, 44(11):1541–1550, 2017.

[300] Sofie Saerens, Johan Thevelein, and Freddy Delvaux. Ethyl ester production during brewery fermentation, a review. *Cerevisia*, 33(2):82–90, 2008.

[301] Brian R. Gibson, Erna Storgårds, Kristoffer Krogerus, and Virve Vidgren. Comparative physiology and fermentation performance of saaz and frohberg lager yeast strains and the parental species saccharomyces eubayanus. *Yeast*, 30(7):255–266, 2013.

[302] J. Peter Gogarten and Jeffrey P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687, 2005.

[303] Claudia Solís-Lemus, Paul Bastide, and Cécile Ané. PhyloNetworks: A Package for Phylogenetic Networks. *Molecular Biology and Evolution*, 34(12):3292–3298, 09 2017.

[304] Hussein A. Hejase and Kevin J. Liu. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, 17(1):422, 2016.

[305] Elizabeth A. Winzeler, Daniel D. Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, Rocio Benito, Jef D. Boeke, Howard Bussey, Angela M. Chu, Carla Connelly, Karen Davis, Fred Dietrich, Sally Whelen Dow, Mohamed El Bakkoury, Françoise Foury, Stephen H. Friend, Erik Gentalen, Guri Giaever, Johannes H. Hegemann, Ted Jones, Michael Laub, Hong Liao, Nicole Liebundguth, David J. Lockhart, Anca Lucau-Danila, Marc Lussier, Nasiha M'Rabet, Patrice Menard, Michael Mittmann, Chai Pai, Corinne Rebischung, Jose L. Revuelta, Linda Riles, Christopher J. Roberts, Petra Ross-MacDonald, Bart Scherens, Michael Snyder, Sharon Sookhai-Mahadeo, Reginald K. Storms, Steeve Véronneau, Marleen Voet, Guido Volckaert, Teresa R. Ward, Robert Wysocki, Grace S. Yen, Kexin Yu, Katja Zimmermann, Peter Philippsen, Mark Johnston, and Ronald W. Davis. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, 1999.

[306] Dan He, Subrata Saha, Richard Finkers, and Laxmi Parida. Efficient algorithms for polyploid haplotype phasing. *BMC Genomics*, 19(2):110, 2018.

[307] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019.

[308] Paul M. Magwene, Ömür Kayıkçı, Joshua A. Granek, Jennifer M. Reininga, Zackary Scholl, and Debra Murray. Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 108(5):1987–1992, 2011.

[309] S R Chambers, N Hunter, E J Louis, and R H Borts. The mismatch repair system reduces meiotic homeologous recombination and stimulates recombination-dependent chromosome loss. *Molecular and Cellular Biology*, 16(11):6110–6120, 1996.

[310] Sara S González, Eladio Barrio, and Amparo Querol. Molecular characterization of new natural hybrids of saccharomyces cerevisiae and s. kudriavzevii in brewing. *Appl. Environ. Microbiol.*, 74(8):2314–2320, 2008.

[311] Caiti S. Smukowski Heil, Christopher G. DeSevo, Dave A. Pai, Cheryl M. Tucker, Margaret L. Hoang, and Maitreya J. Dunham. Loss of Heterozygosity Drives Adaptation in Hybrid Yeast. *Molecular Biology and Evolution*, 34(7):1596–1612, 03 2017.

[312] Caiti S. Smukowski Heil, Christopher R. L. Large, Kira Patterson, Angela Shang-Mei Hickey, Chiann-Ling C. Yeh, and Maitreya J. Dunham. Temperature preference can bias parental genome retention during hybrid evolution. *PLOS Genetics*, 15(9):1–23, 09 2019.

[313] Laura Pérez Través, Christian Ariel Lopes, Eladio Barrio, and Amparo Querol. Study of the stabilization process in saccharomyces intra-and interspecific hybrids in fermentation conditions. 2014.

[314] Zsuzsa Antunovics, Huu-Vang Nguyen, Claude Gaillardin, and Matthias Sipiczki. Gradual genome stabilisation by progressive reduction of the Saccharomyces uvarum genome in an interspecific hybrid with Saccharomyces cerevisiae. *FEMS Yeast Research*, 5(12):1141–1150, 12 2005.

[315] Ksenija Lopandic, Walter P. Pfliegler, Wolfgang Tiefenbrunner, Helmut Gangl, Matthias Sipiczki, and Katja Sterflinger. Genotypic and phenotypic evolution of yeast interspecies hybrids during high-sugar fermentation. *Applied Microbiology and Biotechnology*, 100(14):6331–6343, 2016.

[316] Emil Christian Hansen. Recherches sur la physiologie et la morphologie des ferments alcooliques. v. methodes pour obtenir des cultures pures de saccharomyces et de microorganismes analogous. *Compt. Rend. Trav. Lab. Carlsberg.*, 2:92–105, 1883.

[317] Pierre Gélinas. Mapping early patents on baker's yeast manufacture. *Comprehensive Reviews in Food Science and Food Safety*, 9(5):483–497, 2010.

[318] Ragnhild Scheda and D. Yarrow. Variation in the fermentative pattern of some saccharomyces species. *Archiv für Mikrobiologie*, 61(3):310–316, 1968.

[319] Ian Spencer Hornsey. *A history of beer and brewing*, volume 34. Royal Society of Chemistry, 2003.

[320] F. Mendlik. Some aspects of the scientific development of brewing in holland. *Journal of the Institute of Brewing*, 43(4):294–300, 1937.

[321] Patrick J. Keeling and Jeffrey D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008.

[322] Christopher M. Thomas and Kaare M. Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9):711–721, 2005.

[323] Fernando Racimo, Sriram Sankararaman, Rasmus Nielsen, and Emilia Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, 2015.

[324] Karen E. Nelson, Rebecca A. Clayton, Steven R. Gill, Michelle L. Gwinn, Robert J. Dodson, Daniel H. Haft, Erin K. Hickey, Jeremy D. Peterson, William C. Nelson, Karen A. Ketchum, Lisa McDonald, Teresa R. Utterback, Joel A. Malek, Katja D. Linher, Mina M. Garrett, Ashley M. Stewart, Matthew D. Cotton, Matthew S. Pratt, Cheryl A. Phillips, Delwood Richardson, John Heidelberg, Granger G. Sutton, Robert D. Fleischmann, Jonathan A. Eisen, Owen White, Steven L. Salzberg, Hamilton O. Smith, J. Craig Venter, and Claire M. Fraser. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399(6734):323–329, 1999.

[325] Greger Larson, Keith Dobney, Umberto Albarella, Meiying Fang, Elizabeth Matisoo-Smith, Judith Robins, Stewart Lowden, Heather Finlayson, Tina Brand, Eske Willerslev, Peter Rowley-Conwy, Leif Andersson, and Alan Cooper. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science*, 307(5715):1618–1621, 2005.

[326] Emily Jane McTavish, Jared E. Decker, Robert D. Schnabel, Jeremy F. Taylor, and David M. Hillis. New world cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences*, 110(15):E1398–E1406, 2013.

[327] Rachel Brenchley, Manuel Spannagl, Matthias Pfeifer, Gary L. A. Barker, Rosalinda D'Amore, Alexandra M. Allen, Neil McKenzie, Melissa Kramer, Arnaud Kerhornou, Dan Bolser, Suzanne Kay, Darren Waite, Martin Trick, Ian Bancroft, Yong Gu, Naxin Huo, Ming-Cheng Luo, Sunish Sehgal, Bikram Gill, Sharyar Kianian, Olin Anderson, Paul Kersey, Jan Dvorak, W. Richard McCombie, Anthony Hall, Klaus F. X. Mayer, Keith J. Edwards, Michael W. Bevan, and Neil Hall. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426):705–710, 2012.

[328] G Albert Wu, Simon Prochnik, Jerry Jenkins, Jerome Salse, Uffe Hellsten, Florent Murat, Xavier Perrier, Manuel Ruiz, Simone Scalabrin, Javier Terol, Marco Aurélio Takita, Karine Labadie, Julie Poulain, Arnaud Couloux, Kamel Jabbari, Federica Cattonaro, Cristian Del Fabbro, Sara Pinosio, Andrea Zuccolo, Jarrod Chapman, Jane Grimwood, Francisco R Tadeo, Leandro H Estornell, Juan V Muñoz-Sanz, Victoria Ibanez, Amparo Herrero-Ortega, Pablo Aleza, Julián Pérez-Pérez, Daniel Ramón, Dominique Brunel, François Luro, Chunxian Chen, William G Farmerie, Brian Desany, Chinnappa Kodira, Mohammed Mohiuddin, Tim Harkins, Karin Fredrikson, Paul Burns, Alexandre Lomsadze, Mark Borodovsky, Giuseppe Reforgiato, Juliana Freitas-Astúa, Francis Quetier, Luis Navarro, Mikeal Roose, Patrick Wincker, Jeremy Schmutz, Michele Morgante, Marcos Antonio Machado, Manuel Talon, Olivier Jaillon, Patrick Ollitrault, Frederick Gmitter, and Daniel Rokhsar. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology*, 32(7):656–662, 2014.

[329] JEAN-LUC LEGRAS, DIDIER MERDINOGLU, JEAN-MARIE CORNUET, and FRANCIS KARST. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Molecular Ecology*, 16(10):2091–2102, 2007.

[330] Gianni Liti, David M. Carter, Alan M. Moses, Jonas Warringer, Leopold Parts, Stephen A. James, Robert P. Davey, Ian N. Roberts, Austin Burt, Vassiliki Koufopanou, Isheng J. Tsai, Casey M. Bergman, Douda Bensasson, Michael J. T. O'Kelly, Alexander van Oudenaarden, David B. H. Barton, Elizabeth Bailes, Alex N. Nguyen, Matthew Jones, Michael A. Quail, Ian Goodhead, Sarah Sims, Frances Smith, Anders Blomberg, Richard Durbin, and Edward J. Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341, 2009.

[331] Delphine Sicard and Jean-Luc Legras. Bread, beer and wine: Yeast domestication in the saccharomyces sensu stricto complex. *Comptes Rendus Biologies*, 334(3):229 – 236, 2011. On the trail of domestications, migrations and invasions in agriculture.

[332] Xiaozhou Luo, Michael A. Reiter, Leo D'Espaux, Jeff Wong, Charles M. Denby, Anna Lechner, Yunfeng Zhang, Adrian T. Grzybowski, Simon Harth, Weiyin Lin, Hyunsu Lee, Changhua Yu, John Shin, Kai Deng, Veronica T. Benites, George Wang, Edward E.K. Baidoo, Yan Chen, Ishaan Dev, Christopher J. Petzold, and Jay D. Keasling. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast, 2019.

[333] Siti Hajar [Mohd Azhar], Rahmath Abdulla, Siti Azmah Jambo, Hartinie Marbawi, Jualang Azlan Gansau, Ainol Azifa [Mohd Faik], and Kenneth Francis Rodrigues. Yeasts in sustainable bioethanol production: A review. *Biochemistry and Biophysics Reports*, 10:52 – 61, 2017.

[334] Kristoffer Krogerus, Frederico Magalhães, Virve Vidgren, and Brian Gibson. Novel brewing yeast hybrids: creation and application. *Applied microbiology and biotechnology*, 101(1):65–78, 01 2017.

[335] David Peris, Christian A. Lopes, Armando Arias, and Eladio Barrio. Reconstruction of the evolutionary history of saccharomyces cerevisiae x s. kudriavzevii hybrids based on multilocus sequence analysis. *PLOS ONE*, 7(9):1–14, 09 2012.

[336] Samina Naseeb, Haya Alsammar, Tim Burgis, Ian Donaldson, Norman Knyazev, Christopher Knight, and Daniela Delneri. Whole genome sequencing, de novo assembly and phenotypic profiling for the new budding yeast species saccharomyces jurei. *G3 (Bethesda, Md.)*, 8(9):2967–2977, 08 2018.

[337] Laura G. Macías, Miguel Morard, Christina Toft, and Eladio Barrio. Comparative genomics between saccharomyces kudriavzevii and s. cerevisiae applied to identify mechanisms involved in adaptation. *Frontiers in Genetics*, 10:187, 2019.

[338] Gianni Liti, Alex N Nguyen Ba, Martin Blythe, Carolin A Müller, Anders Bergström, Francisco A Cubillos, Felix Dafhnis-Calas, Shima Khoshraftar, Sunir Malla, Neel Mehta, Cheuk C Siow, Jonas Warringer, Alan M Moses, Edward J Louis, and Conrad A Nieduszynski. High quality de novo sequencing and assembly of the saccharomyces arboricolus genome. *BMC genomics*, 14:69–69, 01 2013.

[339] Arthur R Gorter de Vries, Jack T Pronk, and Jean-Marc G Daran. Lager-brewing yeasts in the era of modern genetics. *FEMS Yeast Research*, 19(7), 09 2019. foz063.

[340] Anthony R. Borneman, Brian A. Desany, David Riches, Jason P. Affourtit, Angus H. Forgan, Isak S. Pretorius, Michael Egholm, and Paul J. Chambers. The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with Saccharomyces cerevisiae and Saccharomyces kudriavzevii origins. *FEMS Yeast Research*, 12(1):88–96, 02 2012.

[341] Huu-Vang Nguyen, Jean-Luc Legras, Cécile Neuvéglise, and Claude Gaillardin. Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of saccharomyces bayanus strains nbrc1948 and cbs380. *PloS one*, 6(10):e25821–e25821, 2011.

[342] Sara S. González, Eladio Barrio, Jürg Gafner, and Amparo Querol. Natural hybrids from Saccharomyces cerevisiae, Saccharomyces bayanus and Saccharomyces kudriavzevii in wine fermentations. *FEMS Yeast Research*, 6(8):1221–1234, 12 2006.

[343] Jennifer R Bellon, Frank Schmid, Dimitra L Capone, Barbara L Dunn, and Paul J Chambers. Introducing a new breed of wine yeast: interspecific hybridisation between a commercial saccharomyces cerevisiae wine yeast and saccharomyces mikatae. *PloS one*, 8(4):e62053–e62053, 04 2013.

[344] Haya F Alsammar, Samina Naseeb, Lorenzo B Brancia, R Tucker Gilman, Ping Wang, and Daniela Delneri. Targeted metagenomics approach to capture the biodiversity of saccharomyces genus in wild environments. *Environmental microbiology reports*, 11(2):206–214, 04 2019.

[345] Tobias Andermann, Alexandre M Fernandes, Urban Olsson, Mats Töpel, Bernard Pfeil, Bengt Oxelman, Alexandre Aleixo, Brant C Faircloth, and Alexandre Antonelli. Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. *Systematic Biology*, 68(1):32–46, 05 2018.

[346] Dominik Schrempf, Bui Quang Minh, Nicola [De Maio], Arndt [von Haeseler], and Carolin Kosiol. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, 407:362 – 370, 2016.

[347] Alastair J. Potts, Terry A. Hedderson, and Guido W. Grimm. Constructing Phylogenies in the Presence Of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. *Systematic Biology*, 63(1):1–16, 10 2013.

[348] C. William Birky. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics*, 144(1):427–437, 1996.

[349] Heidi E.L. Lischer, Laurent Excoffier, and Gerald Heckel. Ignoring Heterozygous Sites Biases Phylogenomic Estimates of Divergence Times: Implications for the Evolutionary History of Microtus Voles. *Molecular Biology and Evolution*, 31(4):817–831, 12 2013.

[350] Brian D. Ondov, Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. Mash screen: high-throughput sequence containment estimation for genome discovery. *Genome Biology*, 20(1):232, 2019.

[351] C. Brown and Luiz Irber. sourmash: a library for minhash sketching of dna. *Journal of Open Source Software*, 1(5):27, 2016.

[352] F. P. Breitwieser, D. N. Baker, and S. L. Salzberg. Krakenuniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, 19(1):198, 2018.

[353] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. In *IN AOFA '07: PROCEEDINGS OF THE 2007 INTERNATIONAL CONFERENCE ON ANALYSIS OF ALGORITHMS*, 2007.

[354] Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A. Cohen, and Mark Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, 2003.

[355] Devin R. Scannell, Oliver A. Zill, Antonis Rokas, Celia Payen, Maitreya J. Dunham, Michael B. Eisen, Jasper Rine, Mark Johnston, and Chris Todd Hittinger. The awesome power of yeast evolutionary genetics: New genome sequences and strain resources for the saccharomyces sensu stricto genus. *G3: Genes, Genomes, Genetics*, 1(1):11–25, 2011.

[356] Ksenija Lopandic, Hakim Tafer, and Katja Sterflinger. Draft genome sequence of the saccharomyces cerevisiae × saccharomyces kudriavzevii ha1836 interspecies hybrid yeast. *Microbiology Resource Announcements*, 6(20), 2018.

[357] Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, 2020/05/12 2019.

[358] Haya Alsammar and Daniela Delneri. An update on the diversity, ecology and biogeography of the Saccharomyces genus. *FEMS Yeast Research*, 20(3), 03 2020. foaa013.

[359] Samina Naseeb, Stephen A James, Haya Alsammar, Christopher J Michaels, Beatrice Gini, Carmen Nueno-Palop, Christopher J Bond, Henry McGhie, Ian N Roberts, and Daniela Delneri. Saccharomyces jurei sp. nov., isolation and genetic identification of a novel yeast species from quercus robur. *International journal of systematic and evolutionary microbiology*, 67(6):2046–2052, 06 2017.

[360] Primrose J. Boynton and Duncan Greig. The ecology and evolution of non-domesticated saccharomyces species. *Yeast*, 31(12):449–462, 2014.

[361] Anthony R. Borneman and Isak S. Pretorius. Genomic insights into the saccharomyces sensu stricto complex. *Genetics*, 199(2):281–291, 2015.

[362] Pavol Sulo, Dana Szabóová, Peter Bielik, Silvia Poláková, Katarína Šoltys, Katarína Jatzová, and Tomáš Szemes. The evolutionary history of Saccharomyces species inferred from completed mitochondrial genomes and revision in the 'yeast mitochondrial genetic code'. *DNA Research*, 24(6):571–583, 06 2017.

[363] Quinn K. Langdon, David Peris, Juan I. Eizaguirre, Dana A. Opulente, Kelly V. Buh, Kayla Sylvester, Martin Jarzyna, María E. Rodríguez, Christian A. Lopes, Diego Libkind, and Chris Todd Hittinger. Postglacial migration shaped the genomic diversity and global distribution of the wild ancestor of lager-brewing hybrids. *PLOS Genetics*, 16(4):1–22, 04 2020.

[364] Alex N Salazar and Thomas Abeel. Approximate, simultaneous comparison of microbial genome architectures via syntenic anchoring of quiver representations. *Bioinformatics*, 34(17):i732–i742, 09 2018.

[365] Samuel V. Angiuoli and Steven L. Salzberg. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3):334–342, 12 2010.

[366] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

[367] Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18(11):1814–1828, 2008.

[368] Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE*, 5(6):1–17, 06 2010.

[369] John R. Tyson, Nigel J. O'Neil, Miten Jain, Hugh E. Olsen, Philip Hieter, and Terrance P. Snutch. Minion-based long-read sequencing and assembly extends the caenorhabditis elegans reference genome. *Genome Research*, 28(2):266–274, 2018.

[370] Cristina G. Ghiurcuta and Bernard M. E. Moret. Evaluating synteny for improved comparative studies. *Bioinformatics*, 30(12):i9–i18, 06 2014.

[371] Sebastian Proost, Jan Fostier, Dieter De Witte, Bart Dhoedt, Piet Demeester, Yves Van de Peer, and Klaas Vandepoele. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research*, 40(2):e11–e11, 11 2011.

[372] Thies Gehrmann and Marcel J T Reinders. Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. *Bioinformatics (Oxford, England)*, 31(21):3437–3444, 11 2015.

[373] Haibao Tang, Matthew D. Bomhoff, Evan Briones, Liangsheng Zhang, James C. Schnable, and Eric Lyons. SynFind: Compiling Syntenic Regions across Any Set of Genomes on Demand. *Genome Biology and Evolution*, 7(12):3286–3298, 11 2015.

[374] Guénola Drillon, Alessandra Carbone, and Gilles Fischer. Synchro: A fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLOS ONE*, 9(3):1–8, 03 2014.

[375] Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J. Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C. Suciu, Sun-Gou Ji, Gülfem Demir, Lizao Li, Berke Ç. Toptaş, Alexey Dolgoborodov, Björn Pollex, Iosif Spulber, Irina Glotova, Péter Kómár, Andrew L. Stachyra, Yilong Li, Milos Popovic, Morten Källberg, Amit Jain, and Deniz Kural. Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, 51(2):354–362, 2019.

[376] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

[377] Ilya Minkin, Anand Patel, Mikhail Kolmogorov, Nikolay Vyahhi, and Son Pham. Sibelia: A scalable and comprehensive synteny block generation tool for closely related microbial genomes, 2013.

[378] Son K. Pham and Pavel A. Pevzner. DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, 26(20):2509–2516, 08 2010.

[379] Andrew S. Warren, James J. Davis, Alice R. Wattam, Dustin Machi, João C. Setubal, and Lenwood S. Heath. Panaconda: Application of pan-synteny graph models to genome content analysis. *bioRxiv*, 2017.

[380] Juan F. Poyatos and Laurence D. Hurst. The determinants of gene order conservation in yeasts. *Genome Biology*, 8(11):R233, 2007.

[381] Arnold Kuzniar, Roeland C. H. J. van Ham, Sándor Pongor, and Jack A.M. Leunissen. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539–551, 2020/02/23 2008.

[382] Walter M. Fitch. Homology: a personal view on some of the problems. *Trends in Genetics*, 16(5):227–231, 2020/02/23 2000.

[383] Chris Duran, David Edwards, and Jacqueline Batley. *Genetic Maps and the Use of Synteny*, pages 41–55. Humana Press, Totowa, NJ, 2009.

[384] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, 06 2015.

[385] MARTIN ERWIG. Inductive graphs and functional graph algorithms. *Journal of Functional Programming*, 11(5):467–492, 2001.

[386] Thomas R. Ioerger, Sunwoo Koo, Eun-Gyu No, Xiaohua Chen, Michelle H. Larsen, William R. Jacobs, Jr., Manormoney Pillay, A. Willem Sturm, and James C. Sacchettini. Genome analysis of multi- and extensively-drug-resistant tuberculosis from kwazulu-natal, south africa. *PLOS ONE*, 4(11):1–9, 11 2009.

[387] Carmen Buchrieser, Philippe Glaser, Christophe Rusniok, Hafed Nedjari, Hélène D'Hauteville, Frank Kunst, Philippe Sansonetti, and Claude Parsot. The virulence plasmid pwr100 and the repertoire of proteins secreted by the type iii secretion apparatus of shigella flexneri. *Molecular Microbiology*, 38(4):760–771, 2000.

[388] Bryan Coburn, Inna Sekirov, and B Brett Finlay. Type iii secretion systems and disease. *Clinical microbiology reviews*, 20(4):535–549, 2007.

[389] Francesc Coll, Ruth McNerney, JoséAfonso Guerra-Assunção, Judith R. Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G. Clark. A robust snp barcode for typing mycobacterium tuberculosis complex strains. *Nature Communications*, 5(1):4812, 2014.

[390] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, 1998.

[391] Christopher R. E. McEvoy, Ruben Cloete, Borna Müller, Anita C. Schürch, Paul D. van Helden, Sebastien Gagneux, Robin M. Warren, and Nicolaas C. Gey van Pittius. Comparative analysis of mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLOS ONE*, 7(4):1–12, 04 2012.

[392] Tanmoy Roychowdhury, Saurav Mandal, and Alok Bhattacharya. Analysis of is6110 insertion sites provide a glimpse into genome evolution of mycobacterium tuberculosis. *Scientific Reports*, 5(1):12567, 2015.

[393] William Klimke, Claire O'Donovan, Owen White, J Rodney Brister, Karen Clark, Boris Fedorov, Ilene Mizrachi, Kim D Pruitt, and Tatiana Tatusova. Solving the problem: genome annotation standards before the data deluge. *Standards in genomic sciences*, 5(1):168–193, 2011.

[394] Maria S Poptsova and J Peter Gogarten. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology*, 156(7):1909–1917, 2010.

[395] Michael. Karas and Franz. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20):2299–2301, 10 1988.

[396] Bogdan Budnik, Ezra Levy, Guillaume Harmange, and Nikolai Slavov. Scope-ms: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology*, 19(1):161, 2018.

[397] Mary E. Norton. Noninvasive prenatal testing to analyze the fetal genome. *Proceedings of the National Academy of Sciences*, 113(50):14173–14175, 2016.

[398] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R. Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh Poudel, and David W. Ussery. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, 15(2):141–161, 2015.

[399] Sophie Zaaijer, Columbia University Ubiquitous Genomics 2015 class, and Yaniv Erlich. Cutting edge: Using mobile sequencers in an academic classroom. *eLife*, 5:e14258, apr 2016.

[400] Yi Zeng and Christopher H. Martin. Oxford nanopore sequencing in a research-based undergraduate course. *bioRxiv*, 2017.

[401] Sarah S Johnson, Elena Zaikova, David S Goerlitz, Yu Bai, and Scott W Tighe. Real-time dna sequencing in the antarctic dry valleys using the oxford nanopore sequencer. *Journal of biomolecular techniques : JBT*, 28(1):2–7, 04 2017.

[402] Thomas Hoenen, Allison Groseth, Kyle Rosenke, Robert J Fischer, Andreas Hoenen, Seth D Judson, Cynthia Martellaro, Darryl Falzarano, Andrea Marzi, R Burke Squires, et al. Nanopore sequencing as a rapidly deployable ebola outbreak tool. *Emerging infectious diseases*, 22(2):331, 2016.

[403] Sarah L. Castro-Wallace, Charles Y. Chiu, Kristen K. John, Sarah E. Stahl, Kathleen H. Rubins, Alexa B. R. McIntyre, Jason P. Dworkin, Mark L. Lupisella, David J. Smith, Douglas J. Botkin, Timothy A. Stephenson, Sissel Juul, Daniel J. Turner, Fernando Izquierdo, Scot Federman, Doug Stryke, Sneha Somasekar, Noah Alexander, Guixia Yu, Christopher E. Mason, and Aaron S. Burton. Nanopore dna sequencing and genome assembly on the international space station. *Scientific Reports*, 7(1):18022, 2017.

[404] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. Deepbinner: Demultiplexing barcoded oxford nanopore reads with deep convolutional neural networks. *PLOS Computational Biology*, 14(11):1–11, 11 2018.

[405] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669, 03 2018.

[406] Martin Hunt, Nishadi De Silva, Thomas D. Otto, Julian Parkhill, Jacqueline A. Keane, and Simon R. Harris. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology*, 16(1):294, 2015.

[407] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 03 2014.

[408] Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, 2016.

[409] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[410] Adam M. Phillippy. New advances in sequence assembly. *Genome Research*, 27(5):xi–xiii, 2017.

[411] Mohamed Ibrahim Abouelhoda and Enno Ohlebusch. Chaining algorithms for multiple genome comparison. *Journal of Discrete Algorithms*, 3(2):321 – 341, 2005. Combinatorial Pattern Matching (CPM) Special Issue.

[412] Todd J. Treangen, Brian D. Ondov, Sergey Koren, and Adam M. Phillippy. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11):524, 2014.

[413] Masa Roller, Vedran Lucić, István Nagy, Tina Perica, and Kristian Vlahovicek. Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic acids research*, 41(19):8842–8852, 10 2013.

[414] Fran Supek. The code of silence: Widespread associations between synonymous codon biases and gene function. *Journal of Molecular Evolution*, 82(1):65–73, 2016.

[415] Biswanath Chowdhury and Gautam Garai. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5):419 – 431, 2017.

[416] Christopher Lee, Catherine Grasso, and Mark F. Sharlow. Multiple sequence alignment using partial order graphs . *Bioinformatics*, 18(3):452–464, 03 2002.

[417] David Zeevi, Tal Korem, Anastasia Godneva, Noam Bar, Alexander Kurilshikov, Maya Lotan-Pompan, Adina Weinberger, Jingyuan Fu, Cisca Wijmenga, Alexandra Zhernakova, and Eran Segal. Structural variation in the gut microbiome associates with host health. *Nature*, 568(7750):43–48, 2019.

Acknowledgments

'*If you want to go fast, go alone. If you want to go far, go together.*'

Indeed, a relevant saying for these past few years.

A Ph.D. is typically a challenging journey, with long-nights and hard deadlines. And it can be particularly difficult if it is in a place far from home. But the friends and colleagues I made throughout my time here has made this journey incredibly fun and exciting. I would like to extend my gratitude to all these people.

First, I would like to thank my promotor, Marcel Reinders, for all the help and support you've given me. It was always encouraging hearing your feedback and advice in my work, despite being responsible for the management and guidance of three entire labs worth-of Ph.D. students. I still remember your comments about the potential culture-clash of my 'salesman-like approach' in my scientific presentations (alas, the American-customs of self-advocacy). Indeed, your pragmatic and honest feedback ultimately helped advance my scientific maturity.

To my supervisor and mentor, Thomas Abeel, this journey would have never been possible without your endless guidance over past several years. It began long ago when we first made acquaintance in Boston; and it is thanks to your eagerness, open-mindedness, and scientific rigour that inspired me to pursue a career in Bioinformatics. Your devotion to scientific communication and detail continues to resonate in my everyday work. I take to heart all the valuable lessons and advice in and outside of the lab.

I would also like to thank the Industrial Microbiology section at TU Delft for all their help and support—making me appreciate the important role of yeasts in our society. A special gratitude to Jean-Marc Daran for all the valuable feedback surrounding yeast biology and guidance in my Ph.D.; Neil Kuipers for the structural support in my projects; Marcel van den Broek for joining me in the 'computational trenches' in our projects; and Pilar de la Torre for the long-read sequencing adventures. Of course, to Nick Brouwers and Arthur Gorter de Vries: Booze-camp will always be ours.

And to the Pattern Recognition and Bioinformatics group: where do I even begin? No, seriously, how should I even *start* to extend my gratitude for all the fond memories?

Should I start with the story about my first *Thursday borrel*, learning about how to 'earn your door'? Or why it is ridiculous to park an SUV? Or let alone, park on a slope?

Maybe I should start with the chip-fiascos? And *The Legend of Bert*? My annual membership to *Reinier de Graaf*? No, the milk didn't go bad? Family-friendly, Christmas-market displays? Dugtrio?

Or what about our dolphin impersonations? Or why we should avoid lab-experiments after 10pm? And while we are it, avoid *Thousand Oaks*?

Or how about the weekly anthem of *Raining Blood*? And *Playing with Fire*? And the late-night U2-cover band?

Maybe I should start with twitter trends?

#RIPfrog
#Suffering
#ThatsMyKapsalon
#Aaaggghhh
#GuldenDraak
#DolphinTrainers
#Pedro
#Kloester
#OneSeventhHeritage
#Wageningen
#O.O.O.G.

In all seriousness, PRB, I will always be grateful for your support and friendship. I hope the above triggered some smiles: you know who you are, and I thank you for all the wonderful memories. I especially send my gratitude to the *usual suspects*: Christian Groß, Laura Cabrera-Quiros, Ekin Gedik, Wouter Kouw, Tom Viering, Alexander Mey, Ahmed Mahfouz, and Marco Loog. As well as the O.O.O.G.s: Christian Groß (again), Stavros Makrodimitris, Christine Anyansi, Tom Mokveld, and Soufiane Mourragui. To Arlin Keo, *happy hardcore* will forever be in my music playlists. To Thies Gehrmann, our fungal research will always matter—and that includes potato as well, Ramin Shirali. To Tamim Abdelaal, you'll always carry the Punch baskeball team on your back. To Sjoerd Huisman, let's not run more late-night experiments again. To Jasper Linthorst, I hope the legend of the American *soccer* player lives on.

Through it all, I could not have gone through this journey without my family. To my better-half, Diana, your love and support have been invaluable—I look forward to our future together. My sisters, Janette and Vanessa, you both always bring joy and optimism to this ever changing world. My cousin, Omar, your work-ethic and resilience is inspiring—you'll always have my support. My sister-from-another-mother, Eva, its been long-days since our first adventures together, don't ever change. And to my parents, Gloria and Alejandro, this degree is dedicated to you.

Curriculum Vitæ

Alex was born on January 28, 1992, in San Jose, California, USA. Son of two outdoor enthusiasts, he frequently hiked and camped throughout the national parks and coastline of California, and became fascinated with the natural sciences around him. But as a native to Silicone Valley, he grew equally interested in technology and innovation.

Motived by his early experiences, Alex pursued a Bachelor of Science at the University of California, Santa Cruz, in Biomolecular Engineering from 2010 to 2015. He became particularly interested in genomics and its potential in personalised medicine for genetic and infectious diseases. As a fellowship-awardee from the National Human Genome Research Institute, Alex joined the Bacterial Genomics group at the Broad Institute of MIT and Harvard from 2013 to 2014, studying how large genomic changes could influence antibiotic-resistance in *Mycobacterium tuberculosis*. His undergraduate experience ultimately motivated him to study algorithm development and application for microbial genomics.

Alex joined the Pattern Recognition and Bioinformatics group at Delft University of Technology as a Ph.D. student from 2015 to 2019. In his research work, he spearheaded efforts for using novel technologies to better understand the genomes and evolution of industrial yeasts. He further developed new, complementary algorithms to unravel hidden patterns in the biology and evolution of these organisms. Towards the end of his Ph.D. research, he became interested in studying the genomes of bacterial viruses, especially in the context of using them as an alternative medicine for treating antibiotic-resistant pathogens.

Alex is currently a Bioinformatic Scientist at SNIPR BIOME, in Copenhagen, Denmark. Together with his colleagues, they aim to develop the next generation of personalised medicine for bacterial and microbiome-related diseases. Alex leads the computational efforts in the company for a wide range of projects, from characterising the genomic landscape of bacterial pathogens and viruses, to modelling and prediction of the interaction between a virus and its host.

List of Publications

9. **Alex N. Salazar**, Franklin L. Nobrega, Christine Anyansi, *et al.*, *An educational guide for nanopore sequencing in the classroom*, PLOS Computational Biology 16(1): e1007314, January 23, 2020.

8. **Alex N. Salazar**, Arthur R. Gorter de Vries, Marcel van den Broek, *et al.*, *Chromosome level assembly and comparative genome analysis confirm lager-brewing yeasts originated from a single hybridization*, BMC Genomics 20, 916, December 2, 2019.

7. **Alex N. Salazar**, Thomas Abeel, *Alpaca: a kmer-based approach for investigating mosaic structures in microbial genomes*, bioRxiv 551234, February 15, 2019.

6. **Alex N. Salazar**, Thomas Abeel, *Approximate, simultaneous comparison of microbial genome architectures via syntenic anchoring of quiver representations*, Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages i732–i742

5. **Alex N. Salazar**, Arthur R. Gorter de Vries, Marcel van den Broek, *et al.*, *Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D*, FEMS Yeast Research, Volume 17, Issue 7, November 2017.

4. Abigail L. Manson, Thomas Abeel, James E. Galagan, Jagadish Chandrabose Sundaramurthi, **Alex N. Salazar** *et al.*, *Mycobacterium tuberculosis Whole Genome Sequences From Southern India Suggest Novel Resistance Mechanisms and the Need for Region-Specific Diagnostics*, Clinical Infectious Diseases, Volume 64, Issue 11, Pages 1494–1501, June 1, 2017.

3. Abigail L. Manson, Keira A. Cohen, Thomas Abeel, Christopher A. Desjardins, Derek T. Armstrong, Clifton E. Barry, III, Jeannette Brand, Sinéad B. Chapman, Sang-Nae Cho, Andrei Gabrielian, James Gomez, Andreea M. Jodals, Moses Joloba, Pontus Jureen, Jong Seok Lee, Lesibana Malinga, Mamoudou Maiga, Dale Nordenberg, Ecaterina Noroc, Elena Romancenco, **Alex N. Salazar** *et al.*, *Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance*, Nature Genetics 49, 395–402, January 16, 2017.

2. Christopher A. Desjardins, Keira A. Cohen, Vanisha Munsamy, Thomas Abeel, Kashmeel Maharaj, Bruce J. Walker, Terrance P. Shea, Deepak V. Almeida, Abigail L. Manson, **Alex Salazar** *et al.*, *Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance*, Nature Genetics 48, 544–551, April 11, 2016.

1. **Alex N. Salazar**, Christopher A. Desjardins, Thomas Abeel, *Normalizing alternate representations of large sequence variants across multiple bacterial genomes*, BMC Bioinformatics 16, A8, January 28, 20120.