

University of Strathclyde, Glasgow, United Kingdom

Strathclyde Institute of Pharmacy and Biomedical Sciences

Genomic and Phenotypic Analysis of Industrially-Relevant Microorganisms

By

Ainsley Dawn McGilvray Beaton

2021

Thesis presented in fulfilment of the requirement for the degree of Doctor of Philosophy

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed: 

Date:29 January 2021

Acknowledgments

Firstly, I would like to thank Nick for being my supervisor since I was even more of a baby scientist than I am now in my undergrad project. I am very glad that something went well enough there that you let me come back for a PhD. The last four years in the lab have been excellent, more to the point the fact that your supervisory style involved frequent trips to the pub for scientific discussion is probably a large contributing factor to me having gotten this far. Further thanks for getting a puppy during my PhD, Tarka always brightens up my week. Also, thanks for preparing me (hopefully I am prepared) for what is both of our dream jobs for me. In all seriousness having a supervisor that usually understands what I am talking about and trusts me anyway even when you don't made a world of difference and I appreciate that as much as the nights propping the bar up in Glasgow, Liverpool, Belfast, and Pisa.

Thanks also to Ingenza, particularly Leonardo and Franck for the opportunity to work with Industry and all of the valuable lessons that have come from that. I have learned a lot from working with both of you and I appreciate all of the catch-up meetings, materials and discussions shared. Also, thanks to Ingenza, IBioIC, and Strathclyde University for funding and extra training.

Thanks to all of the other members of the Micro-group over the years. Especially Alison, Leighton, Talal, and John for their help with all things sequencing. Particular thanks go to all of the members of the group who contributed to the many pre-covid lunches that resulted in tears of laughter, there is nothing else that takes the sting out of ANOTHER failed PCR quite like having all of your friends around. Thanks to Becca and Dave for always staying to the end of the night with me even when none of us should have. Thanks to Liam for giving us all the best stories. Lis for being the kindest human being but also telling everything like it is. Gordon for the cake and gym sessions. Charlie for being the best little lab brother- talking about our failing science, hula mixing, and singing bad 80s music together was the best of Tucker bay times. Walid for being so supportive but mostly for bringing Shawarma. Ally for taking on the English rugby fans with me. Anna for always knowing where to put a full stop and helping me practice my German at 2 am so often. Kirsty and Stuart for being great for a moan and a hill walk. Molly for giving us all such a laugh especially when you weren't trying to and being my favourite snake. This could go on longer than my thesis but thanks also to Jordan, Tiago, Sarah, Gaetan, Adam (for the dog pics), Eilidh, Adrianna, Darren, Elmira, Emily, Jonny, Lily, Robyn, Morgan, Kate, both Pauls, Arnaud and Iain for making the Strathclyde Micro-group such a wonderful, happy and interesting place to work.

Outside of the lab I obviously need to thank my HILLS family, without whom I would be a completely different person who certainly would not have made it this far in life. For managing to simultaneously keep me sane and drive me insane you all deserve a medal. Especially Hannah and Debbie for weekends away, nights spent in Oran Mor, and generally looking after me you guys are the best. Also, Lucy and Sophie for keeping me young and my liver old. Heather for being the first person I remember to ever give me decent uni/life advice when I was 18 and always reminding everyone of two very important things: 1) that it will always be okay and 2) you should wear trainers to the party. I can't name everyone but with you by my side I've been walking on the waves. HTID.

Thanks to my parents for tolerating me being in uni for 9 years, even though you didn't have the first clue what I was doing or why. Mum for allowing me to do what you never had the chance to and keeping me on my toes by making me defend science against whatever had been in a movie, the news, or WhatsApp that week. Dad for showing me how much of a pain in the arse it is when you wait too long to follow your career dreams but doing it anyway and then still moaning when you get the dream job. Elaine for stopping my Dad being too annoying, giving me the ability to afford student life for so long by teaching me effective annoying, and allowing me to afford student life for so long by teaching me the art of effective online shopping from a young age. Though now that I think back, watching you write your master's thesis should have put me off doing this, and I wish you had reminded me of that more.

Gran Nan for probably being the only person in my family who listen long enough to know what I had going on in the lab. Also, for introducing me to Grey's Anatomy and CSI, at far too young an age, that somehow led to me having an interest in Science. Thanks also for giving me my sense of humour, without which I don't think I would have managed. I know that one of your last concerns was me getting this submitted and I wish I would have been able to show you that I had over a 'wee jaunt' and some food of a Saturday afternoon. Oddly, you are probably responsible for at least half of this actually being written due to my unusual productivity on the train back and forth from Hairmyres. To both you and Gran Isobel for showing me what a strong independent woman was way before there was a hashtag for it. Gran Isobel for teaching me how to bake, me and the whole department thank you! For letting me tell you all about my DNA being degraded or my cloning not working even though I was talking complete gibberish. I hope, more than anything that I get to visit you soon and give you some good news about this thesis in person.

To Maya for helping me by talking through all of my scientific problems and checking everything I have ever written. You probably deserve the biggest thanks of all for that as well as the good food, holidays and evenings watching TV on the ceiling.

To Ewan for being the best friend, for the long walks, brunches and too short a time of getting to live together there is no one else I'd want to take thesis writing breaks to watch don't tell the bride with. Also, to Dougs for facilitating the latter half of my thesis writing troll days by feeding and walking me at regular intervals so I didn't die at my computer.

Finally, to Tarka, Scout, Eva, Celyn, Marley, Maisie and Tyeson for all having been sat at my feet while I was doing some sort of PhD related work at some point or another.

Abstract

The field of industrial microbiology is rapidly evolving such that new techniques and methods must be implemented regularly to provide the most efficient industrial processes possible. In collaboration with industrial partners, Ingenza, a combination of high throughput phenotyping and rapid whole-genome sequencing was utilised to provide a greater understanding of industrial production organisms as well as to form a strong platform for further work and analyses. Industrial lineages of *Saccharomyces cerevisiae* for biofuel production and *Streptomyces lavendulae* for mitomycin C production were analysed with whole-genome sequencing to provide further insight into their phenotypic traits and industrial potential. It was found that long-read sequencing of yeast genomes provides a level of assembly completeness and quality difficult to achieve with short-reads alone. This work provides a complete genome sequence for industrial strain Ethanol Red for the first time and shows that it can be used to significantly improve the assemblies of derived production strains. It is also shown that high throughput phenotyping can complement genomic data to provide rapid and functional information about strains in a great variety of growth conditions. This work also provides further evidence that phylogeny and secondary metabolite production often lack correlation as well as functional information as to how best to improve an industrial lineage that presents an unexplained phenotype. Our findings demonstrate that robust genomics pipelines can be utilised to analyse data from personal sequencing devices such as the nanopore and that rapid strain genotyping can be achieved in-house by industry without risking intellectual property integrity. Further, we show that this information can be used to inform experimental protocols that lead to improved production strains.

CHAPTER 1. INTRODUCTION 5

1.1 The Industrial Biotechnology Industry **6**
1.1.1 The Biofuels Industry – an IB growth area. 8
1.1.1.1 Bioethanol 9

1.2 S. cerevisiae **10**
1.2.1 *S. cerevisiae* as a model organism 12
1.2.2 Ethanol Production by *Saccharomyces cerevisiae* 12
1.2.3 Biofuel Producing Yeast 13
1.2.4 Industrial Strains of *Saccharomyces cerevisiae* 14
1.2.4.1 Strain Adaptation for Second Generation Biofuels 14
1.2.5 *S. cerevisiae* genomics 15
1.2.5.1 History and Genealogy 15
1.2.5.2 Genome structure and DNA replication 18
1.2.5.3 *S. cerevisiae* Genome editing tools 19
1.2.6 Whole-genome Sequencing and Analysis of Yeast 22

1.3 *Streptomyces* **22**
1.3.1 *Streptomyces* species in industrial biotechnology 24
1.3.2 *Streptomyces* genomics 25
1.3.2.1 DNA Replication in *Streptomyces* 25
1.3.2.2 Advances in genomics and the effect on *Streptomyces* biotechnology 26

1.4 Advances in Genome Sequencing **26**
1.4.1 Computational Challenges in Genomics 28
1.4.1.1 Base-calling 29
1.4.1.2 Read Error Correction 29
1.4.1.3 Assembly 30
1.4.1.3.1 SPAdes 31
1.4.1.3.2 Long Read Assemblers 33
1.4.1.4 Annotation 35
1.4.1.4.1 RAST 35
1.4.1.4.2 PROKKA 35
1.4.1.4.3 FunGAP 36
1.4.1.4.4 YGAP 36
1.4.1.5 Assembly Quality Assessment 37
1.4.2 Database and Bioinformatics Pipeline Availability 38
1.4.2.1 Read Mapping and Variant Detection 38
1.4.2.2 Genome viewers 39
1.4.2.3 Secondary metabolite cluster prediction 39
1.4.2.4 Whole-Genome Phylogeny Tools 40
1.4.2.5 Phenotypic Analysis 40

1.5 Industrial Strain Phenotyping **40**
1.5.1. Biolog Phenotypic Arrays 41
1.5.2 Relating Phenotype to Genotype 41
1.5.2.1 Genome Wide Association Studies 41

1.6 Aims of the present work **42**

CHAPTER 2 MATERIALS AND METHODS 43

2.1 Strain Acquisition **44**
2.1.1 *S. cerevisiae* 44
2.1.2 Streptomyces 44

2.2 Culture of Microorganisms **44**
2.2.1 Culture Media 44
2.2.2 Antibiotic Concentrations 45
2.1.3 Buffers and Reagents 45

2.3 Plasmids and Primers **45**
2.3.1 Primers 45
2.3.2 Plasmids 45

2.4 *S. cerevisiae* Fermentation **48**

2.5 Whole-genome Sequencing **48**
2.5.1 DNA Extraction 48
2.5.1.1 *S. cerevisiae* short read 48
2.5.1.2 *S. cerevisiae* optimised for long fragment length 48
2.5.1.3 *Streptomyces* short-read DNA Extraction 48
2.5.2 DNA quantification 49
2.5.2.1 Nanodrop 49
2.5.2.2 Qubit Fluorometer 49
2.5.3 Library Preparation 49
2.5.3.1 Illumina Sequencing 49
2.5.3.2 Oxford Nanopore Sequencing 49
2.5.3.2.1 Library Preparation 49
2.5.3.2.2 Native Barcode Ligation 50
2.5.3.2.3 Adapter Ligation and Clean up 50
2.5.3.2.4 Port Priming and Sample Loading 50

2.6 Sequencing data analysis **51**
2.6.1 Sequencing Analysis Workflow overview 51
2.6.1.1 Short reads 51
2.6.1.2 Long reads 52
2.6.1.3 Hybrid assembly 54
2.6.2 Read Assembly 55
2.6.1 De-novo Assembly 55
2.6.1.1 SPAdes 55
2.6.1.2 Canu 55
2.6.2.3 Flye 56
2.6.2.4 Draft Based Scaffolding 56
2.6.3 Assembly Assessment 56
2.6.4 Whole-Genome Alignment and Visualisation 56
2.6.5 Combinatorial Assembly Improvement. 57
2.6.5.1 Pilon 57
2.6.6 Phylogeny inference 57
2.6.7 Secondary metabolite prediction 57
2.6.8 Whole Genome Annotation 57
2.6.10 Whole Genome Sequence Based Comparison 58

2.7 Maintenance of *S. cerevisiae* Plasmids in *E. coli* **58**
2.7.1 Plasmid Confirmation by Restriction Digest 58
2.7.2 Preparation of *E. coli* Competent Cells 59
2.7.3 Transformation of Plasmids to *E. coli* 59
2.7.4 Recovery of plasmid DNA from *E. coli* 59
2.7.5 Restriction Digest of Plasmid 59

2.8 Plasmid Transformation to *S. cerevisiae* **60**
2.8.1 Confirmation of Insertion 60
2.8.1.1 PCR Based Confirmation 60
2.8.1.2 Sequence Based Confirmation 60
2.8.2 Curing of *S. cerevisiae* Integrations 60

2.9 Omnilog: Biolog Phenotypic Microarrays **61**
2.9.1 Plate Preparation 61
2.9.2 Biolog Data analysis 61

CHAPTER 3. ANALYSIS OF INDUSTRIALLY RELEVANT YEAST STRAINS 63

3.1 Introduction **64**
3.1.1 Relevance of Yeast Genome Sequencing to industry 64
3.1.1 Current Reference Strain Availability 64
3.1.3 Aims and Objectives of this Chapter 68

3.2 Results and Discussion **68**
3.2.1 Illumina Sequencing 68
3.2.1.1 Reference Strains 68
3.2.1.1.1 Quality Assessment 68
3.2.1.1.2 SPAdes Assembly 70
3.2.1.2 Industrial Lineage 71
3.2.1.2.1 Quality Assessment 71
3.2.1.2.2 SPAdes Assembly 71
3.2.2 Nanopore Sequencing 73
3.2.2.1 Quality Assessment 73
3.2.2.2 Long Read Assembly 78
Long Read Assembly Quality 80
3.2.2.2.1 Reference-Guided Assembly 82
3.2.2.2.2 Reference-guided Assembly Quality Assessment 84
3.2.3 Whole Genome Annotation 87
3.2.4 Sequence-Based Comparison of Yeast Genomes along an Industrial Lineage 87
3.2.4.1 Comparison of Whole-genome Architecture 87

137
142
3.2.4.2 Single Nucleotide Polymorphisms 150
3.2.1 Heterologous Gene Insertion. 152

3.3 Conclusions **154**

CHAPTER 4. RELATING PHENOTYPE TO GENOTYPE: OMNILOG ASSAYS 155

4.1 Introduction **157**

4.2 Results **157**
4.2.1 Establishing Fermentation conditions in a Parallel Bioreactor System 157
4.2.2 Biolog Phenotypic Arrays 159
4.2.2.1 Omnilog Results 159
4.2.2.1.1 PM3 Nitrogen 159
4.2.2.1.2 PM1 Carbon plate 1 164
167
4.2.2.1.3 PM2 Carbon plate 2 168
170

4.2.2.1.4 PM5 Nutrient Supplements 172
4.2.2.1.5 PM9 osmolytes 176
4.2.2.1.6.PM10 – pH 180
4.2.2.3 Genomics 184

4.3 Conclusions **184**

CHAPTER 5. WHOLE-GENOME SEQUENCING OF AN INDUSTRIAL LINEAGE OF *STREPTOMYCES* 185

5.1 Introduction **186**
5.1.1 *Streptomyces* producing clinically relevant compounds 186
5.1.1.1 Production of the Mitomcyin Group of Compounds 186
5.1.1.2 Mitomycin Mechanism of Action 188
5.1.1.3 Mitomycin Biosynthesis 188
5.1.2 Aims and Objectives 190

5.2 Results **191**
5.2.1 Sequencing output and quality assessment 191
5.2.4 Phylogenetic Placement of Strains of Interest Based on Sequencing Data 198
5.2.5 Assembly correction 203
5.2.6 Assessment of Assembly Completeness 203
5.2.9 Comparison of Average Nucleotide Identity 211
5.2.10 Comparison of Biosynthetic Potential 213
5.2.11 SNPs and Indel Detection 218
5.3 Conclusions and Future work 223

CHAPTER 6. DISCUSSION AND CONCLUSIONS 224

CHAPTER 7. FUTURE WORK 227

CHAPTER 8. REFERENCES 229

APPENDIX 1. 250

APPENDIX 2 263

APPENDIX 3. 270

Introduction to Appendix 3 **271**

Chapter 1. Introduction

1.1 The Industrial Biotechnology Industry

Industrial biotechnology (IB) is a fast-growing and profitable industry that aims to utilise the vast array of biological systems that exist within nature to make commercially useful products. Biotechnology spans many industrial sectors from food provision to healthcare, and this variety of applications is due to the multiplicity of capabilities within the microbial community that can be harnessed. Whilst the concept of using microorganisms for the production of things such as beer and bread is ancient there has been a recent surge in interest in the biotechnology sector (Mattanovich, Sauer and Gasser, 2014; Timmis *et al.*, 2016). This increased interest is partly due to the need to reduce the number of industrial processes that are carried out using finite resources and move towards more sustainable industrial practices, but it is also due to increased understanding of microbial systems.

The first use of biotechnology of any kind was likely in brewing of beer, with reports putting the beginning of brewing history as far back as 6000 years, with the Sumerians being the first to put anything in writing (The Free Library, 2006) . Though this suggests that yeast had been used in this way for even longer still, it wasn't until 'the father of microbiology' Antonie van Leeuwenhoek first reported the existence of what we now know to be microorganisms in the 1680's using his microscope that living creatures of this size were ever considered. It took longer still for these two concepts to be linked, in the late 1830's it was observed that fermentation was caused by self-replicating entities (Schwann, 1836). Since then, the changes to alcoholic fermentation processes have been drastic with most breweries now using pure cultures of their own strain instead of wild yeast (Gibson *et al.*, 2017). Whilst this was independent of any biotechnological knowledge or agenda, this provided a base knowledge and anecdotal understanding that is now utilised in modern biotechnology.

In more recent times the use of microorganisms to produce products for our own benefit has expanded greatly. One of the areas in which microbial production has grown since the beginning of the biotechnological era and continues to be expanded now is in the chemical industry. In 2018, 7% of greenhouse gas emissions were produced by the chemical industry, and this is even after the use of biocatalysts has found a foothold in this sector. Environmental concerns were, however, not the main driver to the rise of industrial biotechnology in this industry, it was the ability of some microbially derived or based processes to be cheaper and faster than their chemical counterparts which makes them truly attractive to industry without requiring the support of subsides or pressure from environmentalists (Gurung *et al.*, 2013; Wilson *et al.*, 2020). For example, some processes that are difficult to complete chemically happen regularly in nature through enzymatic conversions that require far less energy to complete and also have the advantages of regioselectivity and chemoselectivity which can be vital to improving yields and reducing downstream processing times and costs in some chemical processes (Pera *et al.*, 2015; Mu *et al.*, 2020). In 2002, bioprocesses in the chemical industry, excluding biofuels, accounted for $25 billion in revenue and that was only 2% of the chemical industry total (Bachmann *et al.*, 2000). This shows that there has been a relatively long-standing market in this area that generates a significant amount of revenue but still only represents a small fraction of this market. One example of this that can be seen from the use of biocatalysts in big industry is that there is a real need for a genuine economic benefit to exist for the industry in order for these processes to be taken into use. Particularly in cases where

existing processes are long-standing with significant dedicated infrastructure the move to a bioprocess needs to be very well supported and assured in order to be worth the initial capital investment.

Another large-scale use of microbial fermentation is in the manufacture of drugs and biopharmaceuticals including microbially derived products such as antibiotics and anti-cancer drugs from the bacterial genus *Streptomyces*. As with other sectors of industrial biotechnology, the increased ability to use genetic manipulation led to a great deal of growth. This with the already popular bio-based production of interferons and growth hormones occurring in mammalian and plant cell lines were able to be transferred into microbial chassis organisms to make the processes easier and cheaper (Jozala *et al.*, 2016). Recombinant human growth hormone (HGH) is produced in an *E. coli* cell line (Song *et al.*, 2017) with an estimated retail price of $137 M per kg (Puetz and Wurm, 2019). Similarly, Hepatitis B Surface Antigen S Domain is produced by *S. cerevisiae* (Kim *et al.*, 2009) as a vaccine and costs $5.4 B per kg (Puetz and Wurm, 2019). In addition to these types of drugs produced by heterologous expression in microbes there are also a vast number of anticancer, antifungal, and antibacterial molecules produced naturally by microbes namely actinomycetes that will be further discussed later. In all cases genomics has played a large role in the commercialisation of many of these processes.

The exploitation of microbial process is facilitated by the increase in the understanding of microbial metabolic processes. The “genomics era” has allowed for the capabilities of microorganisms to be understood and exploited more widely. Whole-genome sequencing allows for the genetic capacity of microorganisms to be fully realised. With only a small percentage of microbial life being culturable under lab conditions, even with modern culturing techniques and advances in that area, metagenomic sequencing has been imperative in unlocking the full capabilities of the planet’s microorganisms as a genetic resource (Lorenz and Eck 2005). Furthermore, many organisms that can be grown in lab conditions do not produce their full biosynthetic potential when cultured under lab conditions (Palkova, 2004; Okada and Seyedsayamdost, 2017). As such whole-genome sequencing has allowed us to more fully understand the capabilities of whole microbial communities and how to elicit the production of secondary metabolites. In addition to this, modern culturing and sequencing methods have allowed for vast repositories of genetic information to be recorded, stored and made accessible (Gaffney *et al.*, 2020). We now also have a multitude of new molecular biology techniques that allow for an unprecedented variety of microorganisms to be genetically manipulated with relative ease. Techniques such a CRISPR (clustered regularly interspaced short palindromic repeats), and phage integration allow for rapid insertion of genomic material as well as disruption of existing pathways in an organism that may increase the ability of a given strain to produce a molecule of interest. In a cyclic fashion the editing of genomes is also aided by whole-genome sequencing as the information this provides can be used to properly target any changes that are made.

In spite of there being a great deal of competitive industrial biotechnology processes a lot of Current Industrial Biotechnology (CIB) processes fail to keep up with traditional methods of chemical and biofuel production (Erickson, Lutt and Winters, 2016). This is due to several factors including the initial hurdle of capital contribution to move to a bio-based process and

factors such as variable prices of oil and feedstocks that make the market unpredictable. As such Next Generation Industrial Biotechnology is where current interests lie (Yu, Wu and Chen, 2019).This is where low cost, renewable substrates are utilised in energy saving processes to overcome the economic shortcomings of CIB (Yu, Wu and Chen, 2019). The ideal bioprocess would be as or simpler than a chemical process in terms of requirements for product recovery with minimal downstream recovery, be as close to 'carbon neutral' as possible and be part of a circular 'zero waste' bio-economy (Bussey *et al.*, 1995; Seghetta *et al.*, 2016; Venkata Mohan *et al.*, 2016; Butti and Mohan, 2017). In theory this would produce an environmentally conscious and economically viable process but there are many obstacles to achieving this.

1.1.1 The Biofuels Industry – an IB growth area.

Biofuels represent a significant proportion of the industrial biotechnology industry amounting to $136.2 b globally in 2019. This is not least because they seek to replace such a vital part of the petrochemical industry but also because there have been a variety of methods both proposed and in use to achieve an efficient biofuel production method. Biofuels seek to replace and/or augment the use of fossil fuel derived petroleum and diesel to the end of lessening the world's dependence on a finite resource and reducing the environmentally harmful CO_2 emissions produced by the burning of fossil fuels. The biofuel industry is both vast and diverse as, in the search for a biofuel production method that can compete with fossil fuels as well as the other fossil fuel alternatives for price there is a whole host of processes, organisms and feed stocks being utilised. Biofuels include bioethanol, biodiesel and biomethane and the aim is that these bio-produced alternatives will be comparable to the fossil fuel derived versions in terms of octane number, energy content and physical properties in order for them to be utilised without major infrastructure redesign (Peralta-Yahya and Keasling, 2010). The two most commonly used microorganisms in the biofuel producing organism category are yeasts and algae (Francis, 2010; Wijffels and Barbosa, 2010; Singh, Nigam and Murphy, 2011; Buijs, Siewers and Nielsen, 2013; Kwak and Jin, 2017; Spagnuolo, Yaguchi and Blenner, 2019). Algae received attention recently as the proposed solution to all of biotechnology's great problems as it is fast growing, photosynthetic, genetically tractable, phylogenetically diverse and capable of post translational modifications (Chen, Li and Wang, 2019). In many cases it has been found that algae can't quite deliver all that was initially hoped, though there are a number of outfits using algae to produce biofuels as well as a multitude of other commodities. Some of the big advantages to algae over yeast for biofuel production is that they don't need an expensive feedstock since they can produce energy from photosynthesis and that they provide an extra layer of environmental benefits since they consume environmental CO_2 (Lundquist *et al.*, 2010). Algae use 183% of their biomass in CO_2 (Chisti, 2008). Algae grow quickly with a far shorter harvest cycle than plants (<10 days), whilst the majority of algal species require warm climates to grow at this rate, some species are adapted to lower temperatures. *Chlorella* ArM0029, which was first isolated from arctic sea ice can grow relatively well in temperatures as low as $4^{O}C$ (Ahn *et al.*, 2012). Microalgae can grow in both open ponds and photobioreactors, with open ponds currently used for the production of spirulina as a health supplement (Narala *et al.*, 2016; Zhu *et al.*, 2020). The use of open ponds is simpler and more cost effective in that few materials are needed, and no electricity is required for light and/or agitation however, it does mean that algae species able to grow well in conditions that likely contaminants cannot are required to reduce contamination risk (Dahlin *et al.*, 2018;

Ruangsomboon *et al.*, 2020). Algae are good producers of lipids under certain conditions which can be used to produce biodiesel and fatty acid methyl esters (Knothe, 2012; Paul Abishek, Patel and Prem Rajan, 2014). Microalgae are thought to have the potential to produce much larger quantities of biodiesel than other sources such as palm oil and cotton (Schenk *et al.*, 2008; Singh, Nigam and Murphy, 2011). High biomass production is required for algal derived biodiesel to be feasible which would require growth on a much larger scale than exists currently. Microalgae can also be used as a feedstock for other organisms such as *S. cerevisiae* to produce bioethanol via fermentation (Brennan and Owende, 2010; Show *et al.*, 2017) This requires saccharification of the feedstock algae prior to use for digestion of the cell wall which is composed of starch and cellulose. The cyanobacteria, *Synechococcus* sp. strain PCC 7942 has been engineered to produce ethanol aerobically by the addition of heterologous pyruvate decarboxylase (pdc) and alcohol dehydrogenase II (adh) enzymes from *Zymomonas mobilis* (Velmurugan and Incharoensakdi, 2020). Under the expression of a strong native promoter high levels of these enzymes were expressed leading to the successful conversion of solar energy and CO_2 to bioethanol (Deng and Coleman, 1999). Some other advantages to using algae for biofuels are that they don't require the use of arable land or freshwater for growth of themselves or as feedstock and since the strains used for biofuels are not edible there is not an impact on the food chain (Khan, Shin and Kim, 2018). Microalgae have been reported to be able to accrue biomass from solar energy up to 77 g/biomass/m^2/day (Formighieri, Franck and Bassi, 2012; Kirst and Melis, 2014) however, this value can decrease significantly when the scale is increased in either open water or in bioreactors, this is the main limiting factor in the success of algal biofuels currently. In order for any algal based route of alternative fuel production to be viable on a large scale, solutions must be found to increase biomass yield (Singh and Gu, 2010).

1.1.1.1 Bioethanol

The bio-ethanol industry has emerged due to declining fossil fuel resources whilst the demand for transportation fuels still exists. Whilst there are numerous alternatives to fossil fuels for the transportation industry bioethanol is the only one that can be readily used without large changes to engines and infrastructure (Abo *et al.*, 2019). This ease of use is probably the main reason that the bioethanol industry globally was valued at $52.66 billion in 2016 and is predicted to be worth $68.95 billion by 2022 (Market and Market, 2017). Bioethanol also has the advantage of not contributing to the CO_2 emissions released as there is no release of locked carbon (Abo *et al.*, 2019). This means that the CO_2 being released here has only recently been removed from the atmosphere rather than having been locked away for millennia as is the case with fossil fuels. Bioethanol is a diverse option for alternative fuel as it can be produced from a number of starting materials including sugar, maize and lignocellulosic materials (Schenk *et al.*, 2008; Demirbas, 2009; Naik *et al.*, 2010; Ajanovic, 2011; Koizumi, 2014). First-generation biofuels made from feedstocks such as sugar and corn were very popular bioethanol starting materials in the beginning of bioethanol production as they grow fast and produce a good feedstock for the traditional bioethanol producer *Saccharomyces cerevisiae* (Martin, 2010). In the US corn is used widely as a feedstock for biofuels and similarly sugar cane is used in Brazil (Mekonnen *et al.*, 2018). Despite these advantages, such feedstocks are not without their drawbacks and have received criticism mainly on the grounds that their cultivation for bioethanol utilises arable land and water that could otherwise be used to cultivate crops to feed a growing population. These factors prompted the bioethanol industry to look towards

renewable feedstocks and waste streams. Biofuels made from these types of feedstocks that don't require the use of arable land and are composed of waste or by-product from other industries are known as second generation biofuels. This is currently viewed as the most promising type of biofuel (Moon, Kim and Choi, 2012) Lignocellulosic feedstocks include agricultural and forest waste and so using this would eliminate the issues discussed with the use of arable land for bio-ethanol feedstocks. Lignocellulosic biomass (LCB) is the most plentiful source of carbon available to us and is composed of three main components; hemicellulose, lignin and cellulose (Abo *et al.*, 2019). The main challenge that the use of LCB faces is that *S. cerevisiae,* a common micro-organism used in the production of biofuels cannot fully utilise all of the sugars in these types of substrate as they are inaccessible in their natural form (Van Zyl, Bloom and Viktor, 2012). Several methods exist to optimise this type of biomass including pre-treatment to break up the structure of the lignocellulosic fibres, this varies by the specific type of biomass being used but is typically a grinding of the material until fragment size is reduced (Fatma *et al.*, 2018). There are also other methods of micronisation employed including physio-chemical pre-treatment, chemical pre-treatment and biological pre-treatment (Fatma *et al.*, 2018). This part of the process can account for a large proportion of the production costs, but it is essential if the process efficiency is to reach a workable level of feedstock to product conversion rate. After pre-treatment there is a hydrolysis step which de-polymerises the cellulose and hemicellulose into sugars that the yeast can utilise, this can be done either chemically or enzymatically (Nazhad *et al.*, 1995). This makes the use of lignocellulosic biomass problematic, as a significant part of the capital cost of the process is in the pre-treatment of the feedstock (Himmel *et al.*, 2007; Limayem and Ricke, 2012). Corn-mash is one of the candidate feedstocks that second-generation biofuels use. It is reported that using corn-mash as a substrate *S. cerevisiae* can produce 126-130 g/kg of ethanol in a 72-hour fermentation (Devantier, Pedersen and Olsson, 2005). As with many sustainable feedstocks corn-mash is not a pure product, it contains several different carbon sources not all of which can be utilised by native *S. cerevisiae* enzymes. In addition to this, corn-mash can be obtained as a by-product from the fermentation industry where it is the expelled waste from a yeast fermentation in a brewing process. As such, it can vary quite considerably from batch to batch in terms of composition, consistency, and quality. It is estimated that the composition of corn is mainly carbohydrates as a percentage of dry matter with starch making up the majority of that (72%) with only 9.5% and 2.6% of fibre and simple sugars respectively (Ray, 2018). Wild type *S. cerevisiae* does not possess the ability to utilise starch as a carbon source, as such bioprocesses that use corn-mash as feedstock for yeast commonly supplement with glucoamylase and heat stable alpha-amylase to allow for the release of *S. cerevisiae's* preferred carbon source, glucose, from the corn-mash. The addition of enzymes in this manner considerably increases the cost of a bioprocess making it less economically efficient and in turn less attractive to industry.

1.2 S. cerevisiae

The fungus *Saccharomyces cerevisiae* is a single-celled eukaryotic micro-organism. This budding yeast, often referred to as brewer's or baker's yeast is so named for its long-standing history in beer manufacture. Cells divide by a process known as budding where the daughter cells pinch off the mother cell to produce two cells around every 100 minutes under favourable, lab conditions (Herskowitz 1988). This feature makes *S. cerevisiae* distinct from

fission yeasts such as *Schizosaccharomyces pombe* (Hoffman, Wood and Fantes, 2015). Unlike most bacterial genomes, *S. cerevisae* does not have a singular circular chromosome. Instead it has 16 chromosomes and mitochondria combining to form a genome of around 12.1 mbps and over 6000 genes (Fisk *et al.*, 2006). *S. cerevisiae* is known to have two separate mating types denoted by a or α, a haploid a type *S. cerevisiae* can combine with a haploid α type cell to form a diploid cell. When stressed or in nutrient-limiting conditions diploid yeast cells can undergo meiosis and sporulation shown in figure 1.1. which is adapted from Neiman, 2005. This allows one *S. cerevisiae* cell to form four haploid spores that are more resistant to adverse conditions.

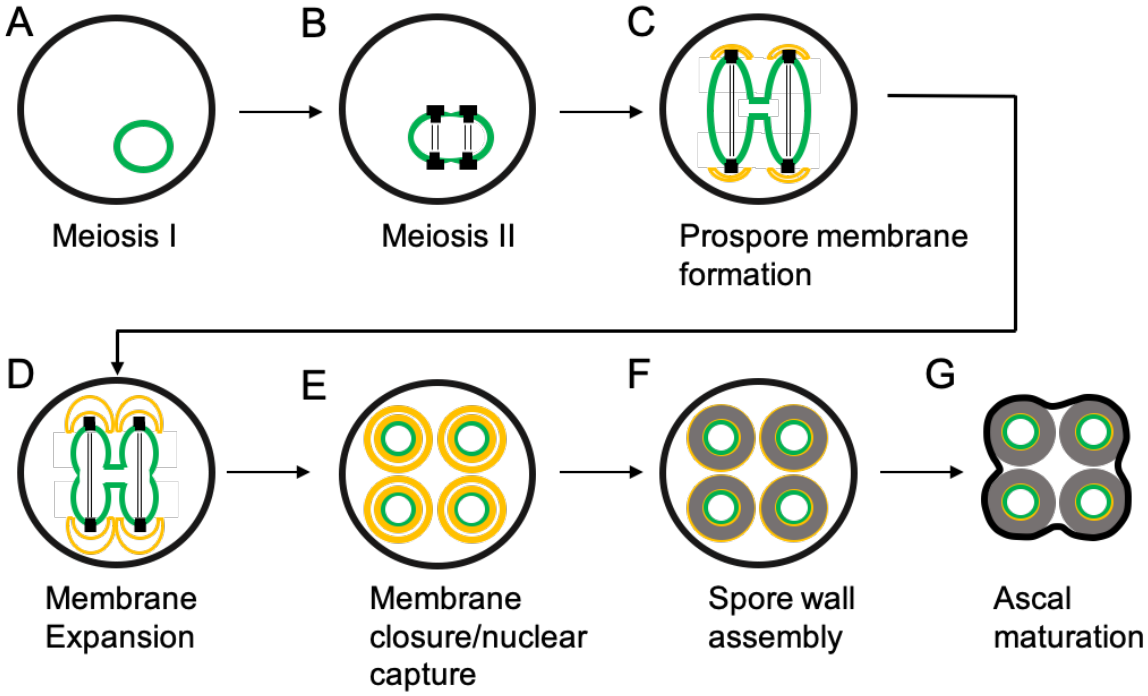


Figure 1.1 The *S. cerevisiae* replication process. A) a cell in a nutrient starved condition in black with the nuclear envelope in green. B) The spindle pore bodies are shown embedded in the nuclear envelope. C) Spindle pore bodies are sites of prospore membrane formation. D) Prospore membrane expands and begins to engulf the haploid nuclei. E) Each prospore membrane closes F) Spore wall membrane synthesis. G) The mother cell collapses and forms the ascus. Adapted from Ascospore Formation in the Yeast *Saccharomyces cerevisiae*. Adapted from Aaron M. Neiman. Microbiology and Molecular Biology Reviews Dec 2005.

1.2.1 *S. cerevisiae* as a model organism

Due to the ability of yeast to grow on agar plates and in liquid medium it can be easily studied in the laboratory. As such they serve as a useful model of eukaryotic genetics and their innate ability to ferment sugars to produce alcohol and CO_2 is a trait that has been widely utilised for centuries. These two factors contribute to make this one of the most well studied microorganisms to date. As a model organism it was instrumental in the early work carried out to understand enzymatic reactions. It was extracts from yeast cells that were used by Eduard Buchner in the early 1900s to show enzymes could be active outside of a cell which led to the first studies of metabolic pathways (Buchner, 1907; Barnett and Lichtenthaler, 2001). The most commonly used lab strain of *S. cerevisiae*, S288c, was selected to be easy to study and maintain. S288c's main precursor strain EM93 was genetically tractable and able to be crossed with other strains which was a vital trait for the initiation of modern understanding of yeast genetics (Mortimer and Johnston, 1986). The crosses made using this organism allowed researchers to understand the functions of many important genes using biochemical mutants. The historical work on *S. cerevisiae* which was largely motivated by its ability to act as a model organism for the study of basic eukaryotic features has led to a wealth of knowledge that can now be utilised. For example, as the first eukaryote to have its whole-genome sequenced there is a great deal of gene characterisation information available in databases such as NCBI (Sayers et al., 2009; NCBI Resource Coordinators, 2017) and Saccharomyces Genome Database (Cherry et al., 1998).

1.2.2 Ethanol Production by *Saccharomyces cerevisiae*

S. cerevisiae, as with many other organisms, can generate ATP via two metabolic processes; respiration and fermentation (Lagunas, 1976; Maskey *et al.*, 2004). Respiration is the default energy production method for most eukaryotes in the presence of oxygen, generating CO_2 and ATP (Pfeiffer, Schuster and Bonhoeffer, 2001; Nilsson and Nielsen, 2016). This process theoretically generates 38 molecules of ATP per glucose molecule, though a lot of factors including the movement of molecules around the cell and variations between different organisms mean that in reality this number is likely to be closer to 18 ATP per glucose molecule (Verduyn *et al.*, 1991; van Gulik and Heijnen, 1995). Fermentation normally occurs only in the absence of oxygen as it generates just 2 ATP per molecule of glucose and is thus considered to be less efficient (Alba-Lois and Segal-kischinevzky, 2010).

A relatively small number of organisms use fermentation as the preferred energy production method, even in the presence of oxygen, this is known as the Crabtree effect (KVAMME, 1958). For an organism to use fermentation over respiration goes against what would be expected, given that it is considerably less energetically efficient. (Postma *et al.*, 1989). There are a few theories to explain this behaviour in yeast, including the rate/yield trade off hypothesis and the whole-genome duplication event (Kellis, Birren and Lander, 2004; Ascencio *et al.*, 2017). The rate/yield trade off hypothesis states that whilst respiration is more energetically efficient in terms of the units of ATP produced per glucose molecule it is slower and so in periods of rapid growth fermentation would be more beneficial to the cell (Pfeiffer, Schuster and Bonhoeffer, 2001).Around 100 million years ago a yeast lineage underwent a whole-genome duplication (WGD) event (Wolfe and Shields, 1997; Kellis, Birren and Lander, 2004; Albertin and Marullo, 2012). Having double the number of chromosomes than was previously required seems like a

drastic change that one might expect to be detrimental to overall fitness, however, this trait has been preserved in many yeast strains and as such it must confer an evolutionary advantage. One of the theories as to the benefits of this event is that genome duplication would have given the potential of the cell to undergo glycolysis at double the rate with double the enzymes (Conant and Wolfe, 2007). Another result of WGD was the eventual generation of two versions of the alcohol dehydrogenase enzyme to give ADHI and ADHII, this gives rise to the "make-accumulate-consume" idea. This is that yeasts would gain a competitive advantage in high sugar niches if it could quickly metabolise the sugars and produce ethanol which would be toxic to many of the competing microorganisms (Piškur *et al.*, 2006). Some of these theories fail to explain the evolutionary pressure that selected for this preference, however Nilsson & Nielsen recently suggested that while respiration is more efficient than fermentation in terms of ATP per molecule of glucose produced it is not more efficient when you consider ATP produced per mass of enzyme (Nilsson and Nielsen, 2016). During periods of rapid growth, when enzyme numbers may be more limited, aerobic fermentation could be the most energetically favourable pathway (Thomson *et al.*, 2005).

Due to the preference that *Saccharomyces cerevisiae* has for fermentation over aerobic respiration it produces ethanol in all conditions, this has been well utilised over the years for bread, wine, and beer production. One of the main advantages to this is that there are constraints placed on any process that needs to be kept either strictly aerobic or anaerobic, but this is one issue that is avoided by using *S. cerevisiae* for ethanol production, these aspects of *S. cerevisiae* metabolic profile makes it the obvious choice for bio-ethanol production.

1.2.3 Biofuel Producing Yeast

While there are several organisms such as non-saccharomyces yeasts, bacteria and algae that are capable of producing ethanol, the obvious choice for an organism for bio-ethanol production is the budding yeast *Saccharomyces cerevisiae.* As well as the reasons already discussed, regarding it being well understood it is also safe to work with being a 'Generally Regarded As Safe' (GRAS) organism (You, Rosenfield and Knipple, 2003). *S. cerevisiae* produces high ethanol yields and has relatively high tolerance to ethanol buildup which makes it perfect for this process.

Ideally the main limiting factor to the ethanol yield of *S. cerevisiae* would be the availability of fermentable sugars, however, the impact of high concentrations of ethanol on cell viability does limit the yield, even though *S. cerevisiae* is more tolerant to ethanol than other microorganisms (Sudha Rani *et al.*, 1996). High ethanol concentrations have an impact on several cellular functions namely cell wall permeability resulting in increased cell death (Kubota *et al.*, 2004; Pina *et al.*, 2004; Georgieva, Skiadas and Ahring, 2007). The link between *S. cerevisiae* ethanol tolerance and membrane composition is well reported; various genes involved in fatty acid, phospholipid, and lipid metabolism and regulation are known to be important in a strain's ability to tolerate ethanol (Ma and Liu, 2010). Screening for mutant libraries has revealed that membrane composition and protein folding were the two most significant factors affecting *S. cerevisiae's* ability to tolerate high ethanol concentrations (Snoek, Verstrepen and Voordeckers, 2016). This knowledge has been used to increase ethanol tolerance in production strains of *S. cerevisiae* by altering the composition of unsaturated fatty acids. *Tni*NPVE, an insect desaturase was inserted into the genome of *S. cerevisiae* resulting in a strain with increased oleic acid to palmitic acid ratio which was shown to confer higher

ethanol tolerance (You, Rosenfield and Knipple, 2003). It has also been shown that the gene *SEC24* is important to *S. cerevisiae's* ability to tolerate ethanol at high temperatures (37^{O}C) and that two amino acid substitutions in this essential gene can impact this ability (Riles and Fay, 2019). Advances such as this, using both directed evolution and targeted mutagenesis have led to the production of *S. cerevisiae* strains that have ethanol production rates and tolerances that are acceptable by the industry under the optimised fermentation conditions.

1.2.4 Industrial Strains of *Saccharomyces cerevisiae*

As there are now many uses for this yeast in various areas of industrial biotechnology, there are a number of strains that have been developed by industry specifically for their particular process. These strains can be generated using screening for naturally favourable qualities as well as via directed evolution or through the insertions of heterologous genes for production of desired proteins. Two of these strains that will be focussed upon in this work are Ethanol Red and Angel (Wallace-Salinas *et al.*, 2015). Ethanol Red is a strain developed for bio-ethanol production that was selected for high ethanol production and tolerance (>18 % v/v) as well as being suitable to high gravity fermentation conditions (Asgher *et al.*, 2010). Angel is also an industrial strain selected for high ethanol tolerance with the same values as Ethanol Red (>18% v/v) (Pandey *et al.*, 2019). Due to both Angel and Ethanol Red being commercial strains, there is little information available on their generation. However, there is a published draft sequence for Ethanol Red which is composed of 216 scaffolds (Wallace-Salinas *et al.*, 2015). There is currently no published sequence for Angel.

1.2.4.1 Strain Adaptation for Second Generation Biofuels

Whilst there have been numerous improvements to bioethanol producing strains, the desire to retain these high ethanol production rates and tolerance levels whilst changing to a more economical feedstock has raised challenges (Eisentraut, 2010; Naik *et al.*, 2010; Sims *et al.*, 2010). The first of these challenges is the requirement for *S. cerevisiae* to utilise a diverse range of carbon sources. This would allow growth and ethanol production on a feedstock such as corn-mash which can be a by-product of other fermentation industries which would lead to the production of 2^{nd} generation bio-ethanol (Chan *et al.*, 2011). This is a feedstock that is economical to obtain and readily available however, it has already had most of the utilisable carbon exhausted. Since it is the by-product of another *S. cerevisiae* fermentation, whilst there may be a great deal of carbon left, everything that *S. cerevisiae* can naturally digest has been depleted. This leads to the requirement to insert heterologous carbon utilisation genes to *S. cerevisiae* to confer the ability to readily utilise all potential carbon sources in a variable feedstock such a corn-mash (Shigechi *et al.*, 2004). In addition to the requirement to break down an expanded range of carbon sources into utilisable sugars there is also the need to be able to transport these carbon sources into the cell. One way to achieve this is by targeted insertion of heterologous genes with a desired function. In order for this strategy to work an extensive knowledge of both general microbial carbon metabolism and *S. cerevisiae* genetics is required. One of the most documented cases of this in the literature is the insertion of amylases to facilitate the utilisation of starch by yeast and limit/negate the need to add amylases to fermentations of starch containing feed stocks. Various combinations of alpha-amylase and promoters have been used to a range of degrees of success. *LKA1* and *LKA2* from *Lipomyces kononenkoae* were inserted on a plasmid under the native PKG1 promoter, when

both of these were inserted in the same strain only low levels of starch utilisation was observed but with fairly stable plasmid maintenance. In the same work SFG1 from *Saccharomycopsis fibuligera* was inserted under its own promoter and terminator region. When expressing this and *LKA1* the strain produced 21 g/l ethanol equal to 0.4 g of ethanol per gram of starch in 120 hours. In all of these cases it was found that starch hydrolysis was still the limiting step of ethanol production (Knox, Du Preez and Kilian, 2004). It has also been found in a separate study where multiple starch degrading enzymes were expressed in *S. cerevisiae* and assessed both individually and in synergy using *A. awamori* GA1, *A. oryzae GlaA* and *AmyI*. It was found that when *GA1* glucoamylase and *AmyI* III α-amylase from *A. awamori* were expressed together there was 61% utilisation of 20 g/L raw starch to produce 5.5 g/L ethanol after 312 h (De Villiers, 2008). The study that investigated the utilisation of raw starch is particularly interesting to this work as it more closely mimics the conditions in corn-mash where the pre-processing may not fully solubilise the feedstock. It is necessary to optimise the carbon utilisation process as this will reduce the costs of the pre-processing and make 2nd generation feedstocks a viable option for ethanol production.

1.2.5 *S. cerevisiae* genomics

1.2.5.1 History and Genealogy

S. cerevisiae S288C is the strain that is most studied in the lab and was the first to be sequenced. It has had a somewhat contrived journey to its current genetic composition that has involved a long series of crosses from various labs resulting in a 'wild type' strain that has been carefully constructed. The strain was generated by Robert Mortimer in the 1950's and its lineage as far back as EM126, FLD, NRRL-210, EM93, Yeast Foam and LK is given in figure 1.2 (Mortimer and Johnston, 1986). This work also showed that most (88%) of the genetic code in S288c is derived from EM93. EM93 is a diploid strain that was originally isolated from rotting figs in Merced, California by Emil Mrak of University of California, Davis in 1938. In 1949 it was reported that EM93 was heterozygous for mating type, galactose fermentation and growth type. The genes involved in these phenotypes were identified by Mortimer et al as *MAT*, *gal2* and *FLO1*. It was also found at this time that EM93 is also heterozygous for *CUP1* which is a metallothionein involved in copper binding and is required for resistance to high concentrations of copper and cadmium (Fogel and Welch, 1982; Lindegren, 2011). It was also determined that EM93 is homozygous for *SUC2* which is an invertase required for hydrolysing sucrose (Carlson, Osmond and Botstein, 1981; Carlson and Botstein, 1982). Interestingly, this work was carried out as a collaboration between R. K. Mortimer at University of California Davis and John R. Johnston at the University of Strathclyde, with these two places having the only two known stocks of strain EM93 at the time of writing the cited paper in 1986 (Mortimer and Johnston, 1986).

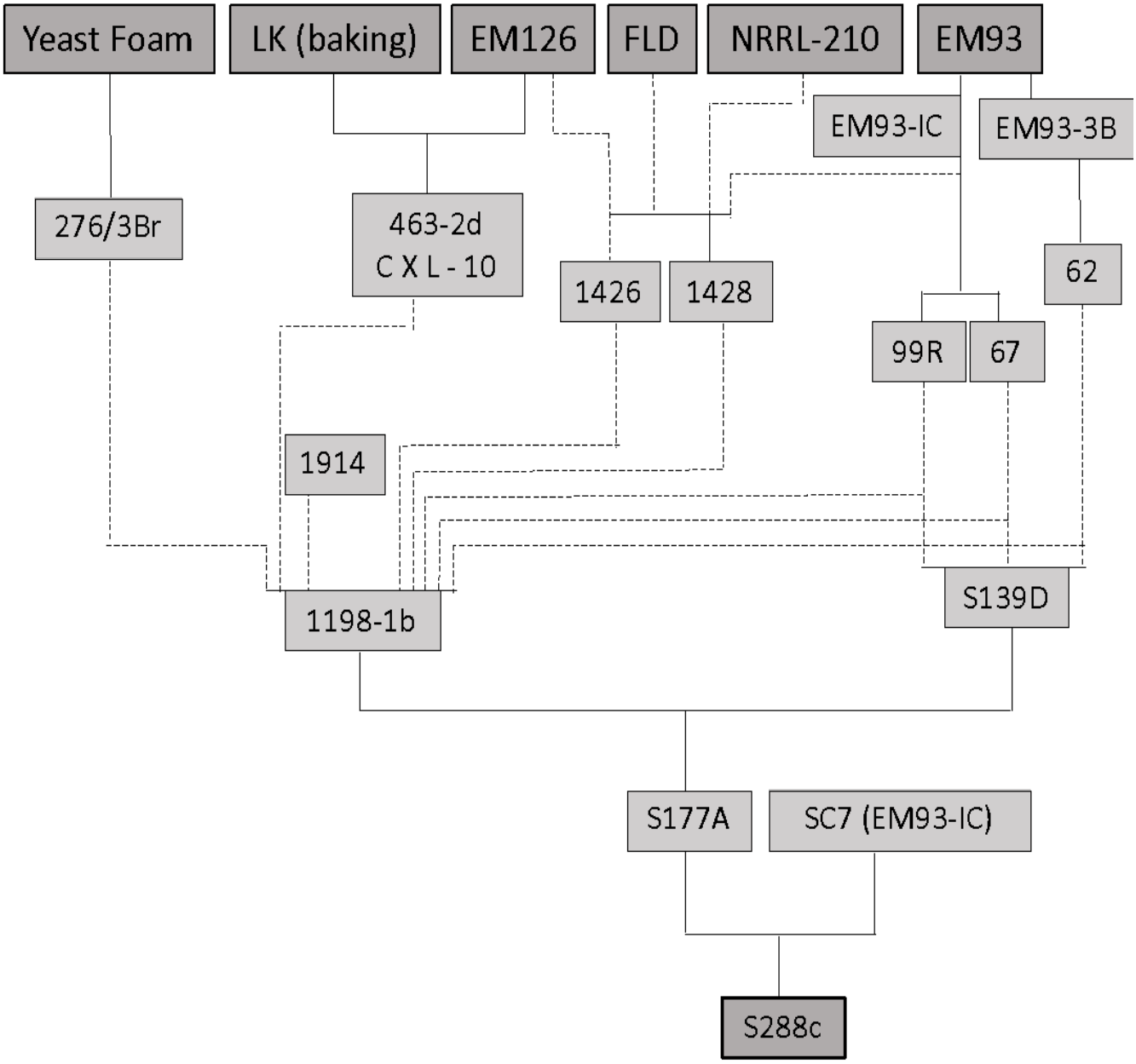


Figure 1.2. Generation of the lab strain S288c. Representation the lineage of strains leading to S288c adapted from figure 2, Mortimer, R. K., and J. R. Johnston. 1986. "Genealogy of Principal Strains of the Yeast Genetic Stock Centre." *Genetics*. Solid lines represent a direct cross, a broken line represents cases where there were also intervening crosses. The strains given here originate from: Mrak: EM93 and EM126; Wickerham:RRL-210; Lindegren.

S. cerevisiae was the first eukaryotic genome to be fully sequenced and was released in 1996 and as of June 2020 there were 825 *S. cerevisiae* genomes available in GenBank (Engel *et al.*, 2014). This first sequencing effort was a global collaboration between researchers in 19 countries and 94 laboratories (Goffeau *et al.*, 1996) resulting in a novel collaborative project (Johnston, 1996). Due to the obvious constraints of this type of collaborative project, the data that made up the initial sequence of S288c was actually composed of reads from several different strains that were considered isogenic to S288c (Engel *et al.*, 2014). This feat allowed a great deal of further work and understanding of eukaryotic genomics but was repeated using modern sequencing technologies and in a standalone experiment from a single colony of S288c in 2010. It is this sequence that is most commonly used today and also that which will be used for reference in this work. Advances in high-throughput sequencing technology means that a single researcher can sequence and analyse the genome of a yeast strain in a relatively short time at significantly lower cost. Figure 1.3 shows a graph published by the NIH showing how rapidly the cost of sequencing has declined in the last two decades, significantly surpassing the expectation set out by Moore's Law and enabling the rapid expansion in utilisation of whole-genome sequencing that we are seeing currently (Wetterstrand KA, no date; Vardi, 2005).

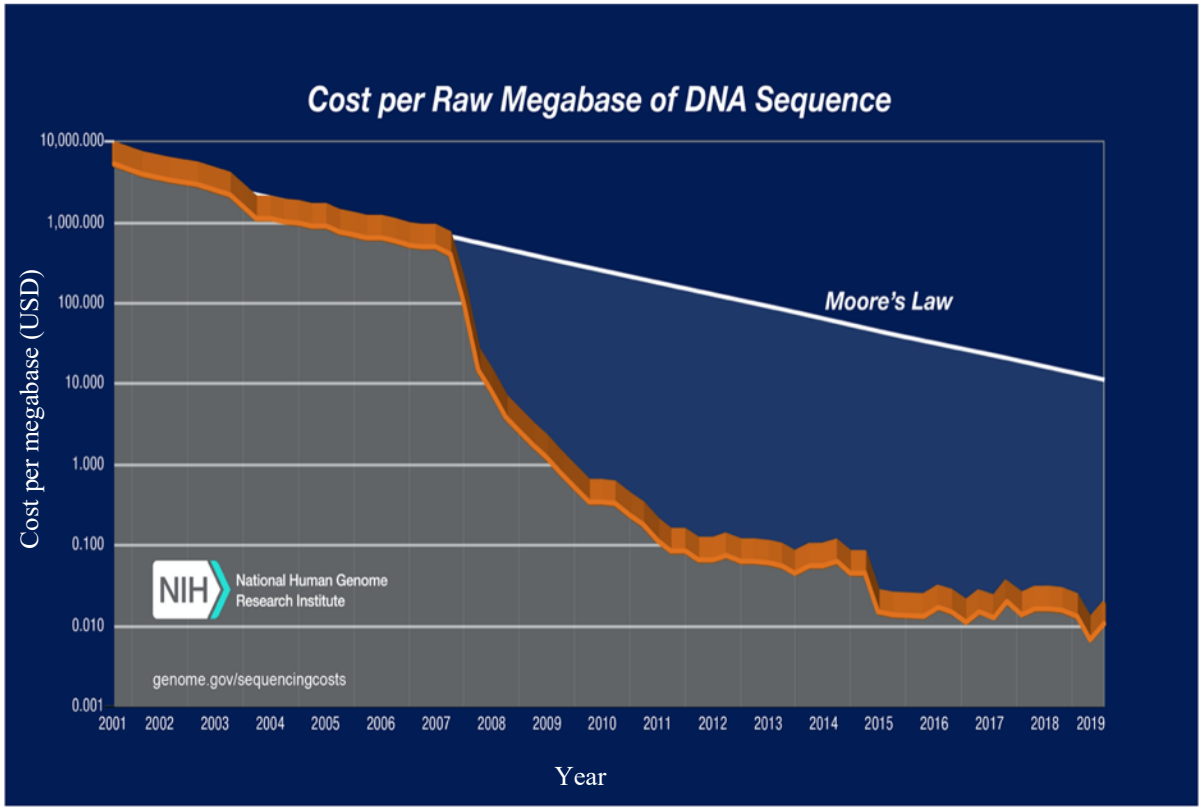


Figure 1.3. Sequencing Cost compared to Moore's Law. Cost of sequence per megabase compared to Moore's Law taken from Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 13/06/2020

1.2.5.2 Genome structure and DNA replication

The 16 chromosomes of *S. cerevisiae* are thought to have arisen after an event called the whole-genome duplication (WGD), prior to which *S. cerevisiae* had only 8 chromosomes. Whilst *S. cerevisiae* is a eukaryote and its genome is structured as such, in a lot of aspects it is simpler and more similar to bacteria when compared to mammalian genomes (Foury, 1997; Duina, Miller and Keeney, 2014). When one compares the density of protein-coding regions in the *S. cerevisiae* genomes to that of a human genome there is approximately fifty times greater gene density in the *S. cerevisiae* genome (Duina, Miller and Keeney, 2014). This can be partially explained by the fact that introns in yeast genomes are fairly rare, only around 5% of *S. cerevisiae* genes are known to contain an intron (Parenteau *et al.*, 2008). Noncoding genes in the genome are largely those that are involved in the transcription and translation as well as regulation of these processes. There are also a number of non-coding RNAs, which have fairly recently been further understood thanks to technological advances including next generation sequencing technologies (Nagalakshmi *et al.*, 2008; Vera and Dowell, 2016). It was originally thought that the un-transcribed portion of eukaryotic genomes was larger than we now know it to be. Non-coding RNAs, especially those located near coding sequences, play a role in transcriptional regulation (Wu, Delneri and O'Keefe, 2012). Non-coding RNAs can be subdivided into several different types including CUTs (Cryptic Unstable Transcripts) (Vera and Dowell, 2016) MUTs (Meiotic Unannotated Transcripts) (Lardenois *et al.*, 2011) and Yeast TERRAs (Telomeric Repeat associated cryptic ncRNAs) (Isken and Maquat, 2009).

The somewhat complex genomics and replication abilities of *S. cerevisiae* mean that there are several features that allow for the correct replication and movement of all 16 chromosomes during either mitosis or meiosis. Each chromosome has several origins of replication, typically occurring every 40 kb or less (Fanning, Knippers and Winnacker, 2012). These were first coined autonomously replicating sequences (ARSs) as they were found to confer the ability of non-replicative plasmid to replicate. It is now known that not all ARSs are 'real' origins of replication, since only a portion of them naturally initiate replication.

Chromosome segregation during replication requires kinetochore microtubules to bind to the chromosome at the centromere. Yeast centromeres are different to the normal eukaryotic centromeres with them being shorter and lacking heterochromatin (Feldmann *et al.*, 1994). As is necessary in all linear pieces of genetic material, each of the *S. cerevisiae* chromosomes is flanked by telomeres, which are repetitive non-coding sequences that protect the integrity of the chromosomes (Teixeira, 2013). Telomeres in *S. cerevisiae* have been studied as a model organism for other eukaryotes, thus making them relatively well understood. *S. cerevisiae* telomeres end in a single stranded g-tail of around 15 bp, but the length of this does vary throughout the cell cycle (Wellinger, Wolf and Zakian, 1993). Immediately preceding the g-tail is the 200-400 bp region of telomeric repeat sequences which are heterogeneous in *S. cerevisiae,* this is quite unusual for telomere sequences but can be useful experimentally as newly synthesised telomere sequence can be distinguished from old. Since *S. cerevisiae* contains only one telomerase RNA gene this is caused by a number of mechanisms including only a section of the mRNA being used or by binding at multiple registers in the transcript (Förstemann and Lingner, 2001). Before the telomeric repeat come the Y elements which are 5.2 or 6.7 kb in length and are not always present. Finally, the X-element which is always

present and is itself composed of the STR (Sub-telomeric repeated elements) and the core X (Wellinger and Zakian, 2012)

1.2.5.3 *S. cerevisiae* Genome editing tools

Homologous recombination was commonly used to insert sequences to the yeast chromosome from a circular plasmid until the 1980s. Whilst this works it had low transformation efficiency and could result in multiple insertions depending on the structure of the plasmid and it's regions of identity to the genome. In 1981 it was reported that a double strand break (DSB) increases the efficiency of the transformation and that the break is always repaired during integration. This also carried the benefit that the likelihood of the homologous region containing the DSB increased, which improved the ability to better target the desired integration (Orr-Weaver, Szostak and Rothstein, 1981). This technique, which is still used today, results in a repeat of the homologous region after chromosome insertion such as the loss of the plasmid backbone and is termed "ends-in" recombination (Thaler, 1988). Further to this several advances were made that led to the use of linear fragments of DNA that contained a selectable marker with homologous flanking regions that would result in stable chromosomal integration (Rothstein, 1983). This is termed "ends-out' recombination and has been used extensively due to its simplicity (Thaler, 1988; Miné-Hattab and Rothstein, 2013).

S. cerevisiae has been the target of genome editing technologies since the 1980s and as such there are a wide variety of methods available for use in this organism. Initial methods of genome editing in *S. cerevisiae* were based on the so called meganucleases, that were first known to be highly specific restriction enzymes responsible for introducing double strand breaks required for the self-splicing of *S. cerevisiae* mitochondrial rDNA introns (Colleaux *et al.*, 1988). Site-specificity is one of the required traits of genome editing tools and soon meganucleases were being utilised that were galactose inducible (I-SceI) (Plessis *et al.*, 1992). This was further developed by Storici et al in 2003 to produce a tool that was both counter-selectable and inducible using the *delitto perfetto* system-based CORE cassette to produce the CORE-I-SceI cassette (Storici *et al.*, 2003). This was later further improved by replacing URA3 marker in the cassette with a herpes simplex virus thymidine kinase which is not typically found in fungal genomes and thus removed the step of having to delete the native URA3 before the CORE-I-SceI could be used (Alexander, Doering and Hittinger, 2014). This satisfied a lot of the needs of genome editing tools, however, as native meganucleases are limited in where their target sites are, methods which utilise these must be preceded by a step to insert the recognition site in its desired location. To get around this odd paradox, engineered meganucleases have been created to change their binding site, but they are difficult to produce, making the process less appealing for rapid genetic modification (Alexander, 2018).

Transcription activator-like effectors (TALEs), which were first reported in 2007, are proteins that naturally alter the gene expression in plant hosts when expressed by pathogenic plant bacteria (Sanjana *et al.*, 2012). TALEs have near repeats of amino acids which dictate their nucleic acid recognition site (Kay *et al.*, 2007; Römer *et al.*, 2007). When these are combined with the non-specific nuclease *FokI* the result is a highly specific site-cleavage system termed TALENs (TALE nucleases) (Li *et al.*, 2011; Aouida *et al.*, 2014). These are much easier to design than site specific meganucleases, though they can be difficult to assemble due to their long repeats (Alexander, 2018).

The CRISPR-Cas systems of gene editing have revolutionised genome editing. Despite there being several published CRISPR systems for use in *S. cerevisiae* the impact of this development on the yeast community has been much smaller than with other organisms (Alexander, 2018). In addition, technology such as this can encounter intellectual property issues when used in industry which can make them unappealing to business. Reasons for this include the fact that it has been found that different strains of *S. cerevisiae* respond differently to each cas system which limits applicability and the fact that before CRISPR there was already traditional methods that were not significantly outperformed by CRISPR.One of the benefits that CRISPR has in yeast that other mechanisms do not is the ability for it to be used to make previously unprecedented numbers of simultaneous mutations (Walter *et al.*, 2019).

In yeast, as with other eukaryotes, when Cas9 and guide RNA (gRNA) are expressed in a yeast cell the Cas9 introduces a double stranded break at the target site of the gRNA. gRNA is an RNA sequence that binds to the DNA region of interest and guides the editing enzyme to this specific DNA site. This break is repaired by non-homologous end-joining of homologous recombination (Jasin and Haber, 2016; Brinkman *et al.*, 2018). By adding a DNA repair template that will be used in homologous recombination it is possible to achieve desired recombination. The main advantage to CRISPR/Cas based gene editing in yeast is that there is no need for the use of a selectable marker so long as there is efficient cutting by Cas9, the double stranded break that is generated will serve for negative selection (Stovicek, Holkenbrink and Borodina, 2017). Other genetic tools available for yeast genetic modification include plasmid insertion with the gene of interest flanked by regions homologous to the genome to facilitate homologous recombination of the gene of interest into the genome. This will typically include a selectable marker such as an antibiotic resistance cassette. In the case of industrial microorganisms, it is important that strains can be generated that do not include any additional antibiotic resistance genes before being used in biotechnology. As such methods are also required to excise these cassettes. The LoxP/Cre recombinase system is commonly used to remove specific genomic regions (Araki *et al.*, 1995).

The Cre-LoxP system is one that is widely used in *Saccharomyces* genetics as it utilises the organism's natural ability to readily undergo homologous recombination and can result in a markerless insert (Banerjee and Biswas, 2008). It also has the benefit of working in non-*Saccharomyces* yeasts such as *Schizosaccharomyces pombe* (Gueldener, 2002; Hentges *et al.*, 2005; Teresa Avelar *et al.*, 2013) and *Kluyveromyces marxianus* (Ribeiro *et al.*, 2007) as well as other model organisms including mice (Gu *et al.*, 1994) and zebrafish (Dong, Liu and Shao, 2010). The Cre-LoxP system comprises the 34 bp *lox* recognition site and the Cre recombinase enzyme (Mack *et al.*, 1992). Depending on the orientation of the *lox* site deletion, inversion and translocation can all be achieved as shown in figure 1.4. Using this method, a selectable marker that is contained between the *lox* sites can be removed, leaving only the *lox* scar site.

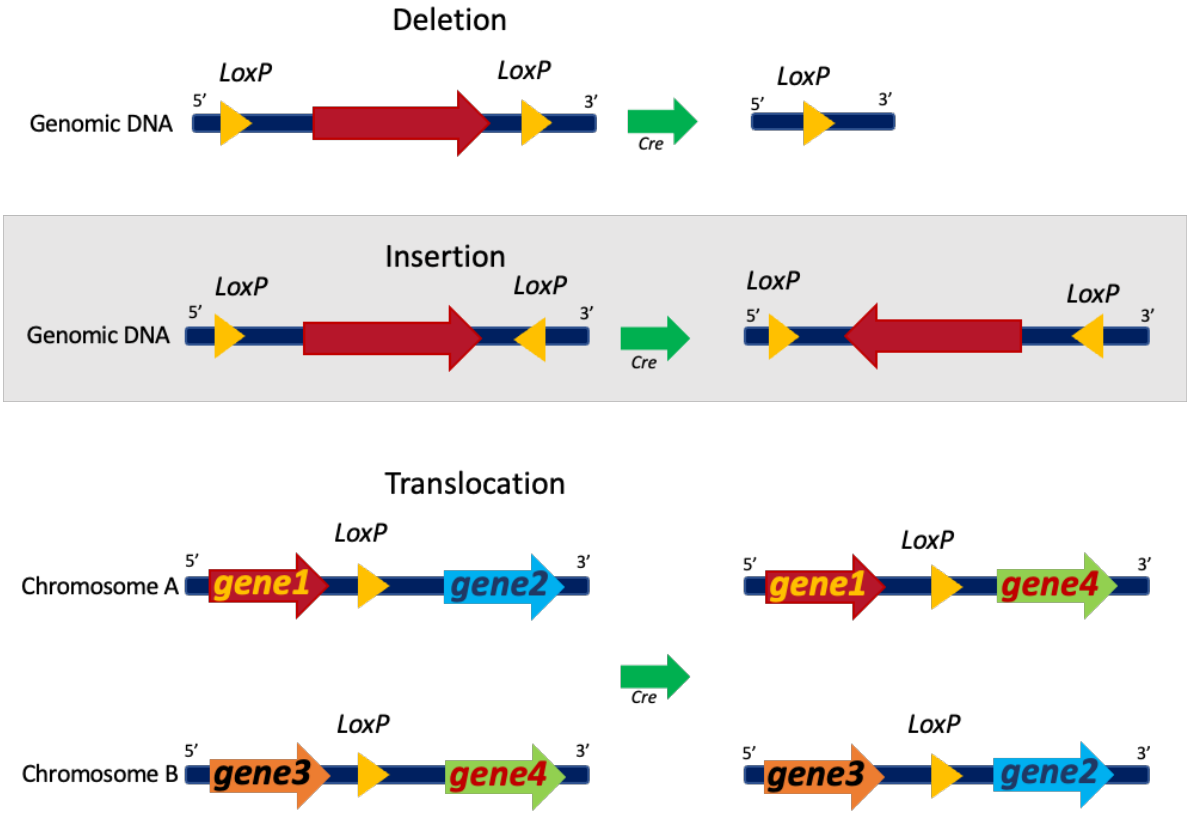


Figure 1.4 *S. cerevisiae* cloning mechanism. Positioning and orientation of the LoxP region in relation to the insertion sequences for deletion, inversion, and translocation.

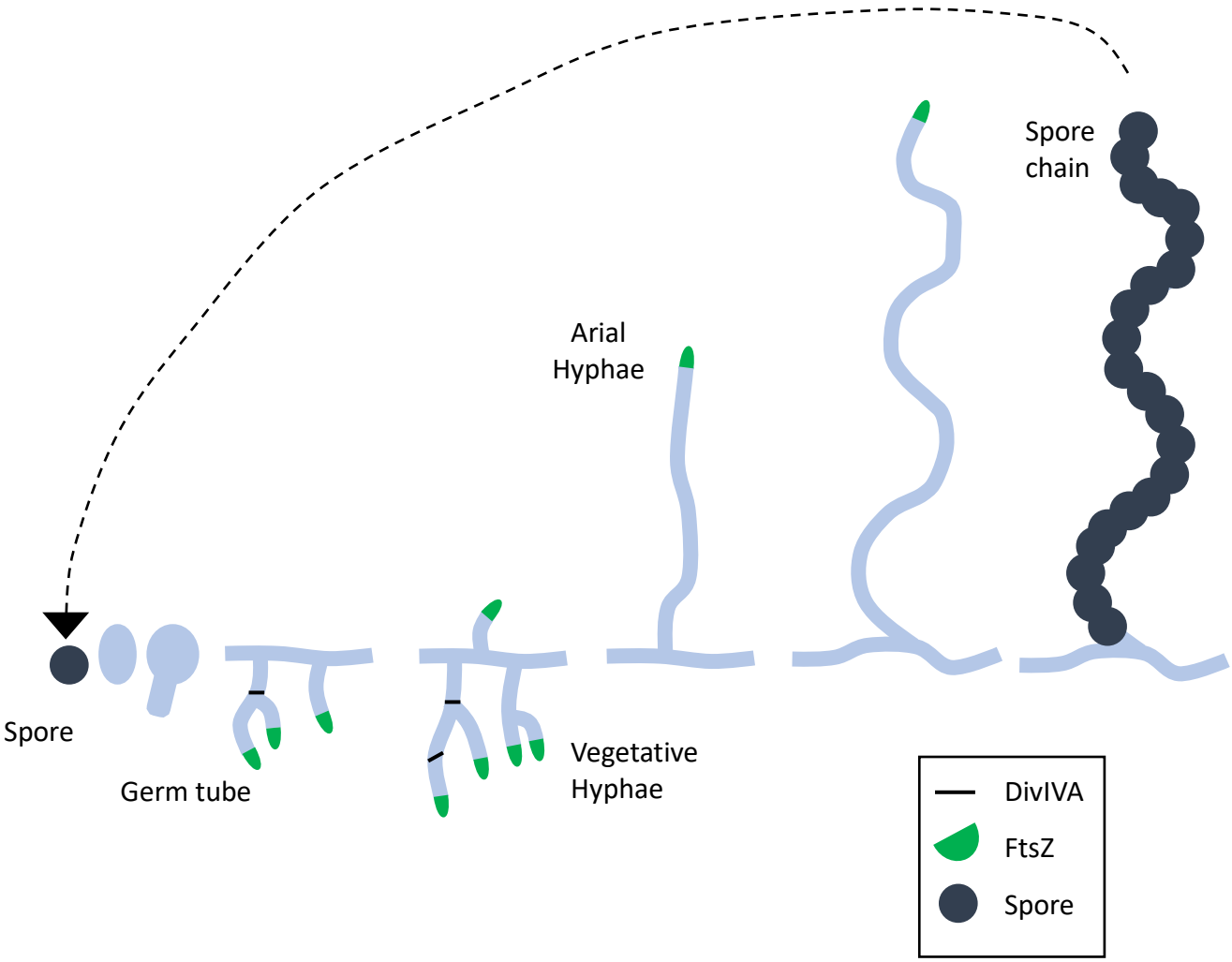
1.2.6 Whole-genome Sequencing and Analysis of Yeast

Since genomes govern a strain's ability to produce ethanol as well as the organism's stress responses which impact on overall yield, knowledge of the genome helps inform how best the organism could be improved to meet the industrial requirements. The first *S. cerevisiae* strain sequenced was the reference strain S288c, this was released in 1996 after its whole-genome sequence had been elucidated by several labs around the world in a consortium effort (Johnston *et al.*, 1994; Bussey *et al.*, 1995, 1997; Galibert *et al.*, 1996; Goffeau *et al.*, 1996; Bowman *et al.*, 1997; Jacq *et al.*, 1997; Philippsen *et al.*, 1997; Churcher *et al.*, 1997; Dietrich *et al.*, 1997; H Tettelin *et al.*, 1997; H. Tettelin *et al.*, 1997; Brennan and Owende, 2010). Since then, this same strain has been re-sequenced using more modern sequencing chemistries and analysis software. There have also been several other strains of *S. cerevisiae* sequenced and deposited to public databases such as NCBI (National Centre for Biotechnological Information) and SGD (Saccharomyces Genome Database) (Fisk *et al.*, 2006; NCBI Resource Coordinators, 2017). Of particular interest is the sequence of the industrial strain Ethanol Red, which was sequenced to draft level in 218 contigs (Wallace-Salinas *et al.*, 2015). The increased rate of sequencing of all organisms has led to a vast increase in the number of platforms available for the analysis of sequencing reads which allows for more accurate assemblies to be generated. Since the advent of genome sequencing different pipelines for analysis have been developed and the algorithms are consistently being improved, furthermore there are pipelines that have been tested and optimised to different organisms and sequencing platforms.

1.3 *Streptomyces*

Another example of an organism widely used in IB fermentations is the bacterial genus *Streptomyces.* Part of the phylum Actinobacteria, they are Gram-positive soil organisms often found in both terrestrial and marine environments. They have multi-stage life cycles and complex genomics which makes them capable of surviving in a wide variety of niches and exhibiting a great wealth of biosynthetic potential (Demain, 1999; Fisk *et al.*, 2006; Lim *et al.*, 2009; Heitefuss, 2010; Kamjam *et al.*, 2017). *Streptomyces* are the most well studied genus of the actinobacteria and are believed to be the most abundant (Olanrewaju and Babalola 2019). The complex life cycle of *Streptomyces* starts and ends with spores that can germinate under favourable conditions to form a vegetative mycelium by tip extension and hyphal branching (Chater, 2016; Koepff *et al.*, 2018). When nutrients become more limiting aerial hyphae are erected and, from these, spores are formed. The activation of *bld* genes during aerial hyphae formation is associated with the onset of production of secondary metabolites. The life cycle of *Streptomyces* species is depicted in figure 1.5. *Streptomyces* can be found in a wide variety of niches, from garden soil to arctic tundra, marine sediment and desert soils (Zvyagintsev, 1994; Okoro *et al.*, 2009; Barke *et al.*, 2010; Heitefuss, 2010; Seipke *et al.*, 2012; Dalisay *et al.*, 2013; Oh *et al.*, 2019). They are also able to survive adverse environmental conditions by sporulating during these times and only germinating once conditions become more favourable for growth. They often grow in environments where there are limited resources and a great deal of competition for these resources from both similar and distinct organisms (Little *et al.*, 2008; Vaz Jauri *et al.*, 2013). This is thought to be the main reason that *Streptomyces* species

produce such a vast number and variety of secondary metabolites, including a variety of economically relevant compounds such as antibiotics, anti-fungals and anti-cancer compounds (Lim *et al.*, 2009; Barke *et al.*, 2010; de Lima Procópio *et al.*, 2012).



1.3.1 *Streptomyces* species in industrial biotechnology

The abundance of secondary metabolites produced by *Streptomyces* is a large contributor to the reason they have been so widely studied. After the discovery of penicillin from the fungus *Penicillium chrysogenum* in 1941, which functions by inhibiting cell wall biosynthesis, there began an 'antibiotic revolution' where several new antimicrobial compounds were found, largely from *Streptomyces* species. The first was streptomycin in 1944 closely followed by most of the antibiotics in clinical use today. Some *Streptomycetes* can produce 30-50 secondary metabolites (Katz and Baltz, 2016). There are a number of theories surrounding the evolutionary advantage for evolving to produce such complex molecules, but the most promising are those that discuss the advantages to the cells that produce molecules that inhibit the growth of other organisms or that these molecules could be used for signalling between cells (Challis and Hopwood, 2003; Yim, Wang and Davies, 2007). *Streptomyce*s produce more than a third of all clinically used anti-bacterial compounds currently used and also a great number of anti-cancer drugs and anti-fungal compounds (Lim *et al.*, 2009; Heitefuss, 2010; Seipke *et al.*, 2012). After streptomycin was isolated from *Streptomyces griseus* in 1944 the 'golden age' of antibiotic discovery began. The initial surge in antibiotic discovery was possible because this was the first time that people were looking in soil for bacteria that had antimicrobial properties and this search found such things as vancomycin from *Streptomyces orientalis* (Griffith, 1984) which inhibits cell wall biosyntheis, streptomycin from *Streptomyces griseus* (Rake and Donovick, 1946) and neomycin from *Streptomyces fradiae* (Waksman and Lechevalier, 1949) which both function by inhibiting protein synthesis. With more antibiotics being discovered and used worldwide the issue of antibiotic resistance was now apparent and some 30 years after penicillin was first used clavulanic acid produced by *Streptomyces clavurigerous* was identified as a beta-lactamase inhibitor that would impede the mechanism of resistance to penicillins (Saudagar, Survase and Singhal, 2008; Tapiero, 2008). Clavulanic acid was first isolated and tested as an antimicrobial itself but it was shown to be ineffective, though in combination with penicillins is it very effective by irreversibly binding to beta-lactamase inhibitors to prevent resistance and is still in use today (Hale, 2019). Eventually the discovery of new antibiotics slowed and those that were found had issues such as cross resistance with other antibiotics, short half-life or toxicity in human cells and so the golden age of antibiotic discovery ended (Silver, 2011). Since then many other techniques have been used such as chemical synthesis of antimicrobials, genomics-based antibiotic detection and looking for novel chemistry in understudied environments (Subramani and Aalbersberg, 2013; Kamjam *et al.*, 2017; Subramani and Sipkema, 2019). Despite this the industrial interest does not lie with antibiotic discovery due to its suboptimal business potential (Hu, 2018; Langreth, 2019). Instead, industry is more interested in products that can be used to treat other human diseases such as cancer. *Streptomyces* also produce several anticancer molecules such as Dactinomycin from *Streptomyces parvullus* (Agrawal, 2007), Bleomycin from *Streptomyces verticillus* (Vig and Lewis, 1978) and mitomycin C from *Streptomyces caespitosus* (Abraham *et al.*, 2006). One issue facing the use of any natural product from *Streptomyces* is eliciting its production under lab conditions or understanding its biosynthesis well enough to make the molecule without the original production organisms. One field that has advanced this kind of work a great deal is genomics (Moir *et al.*, 1999).

1.3.2 *Streptomyces* genomics

Streptomyces have complex genomes, unlike most bacteria they have large linear chromosomes ~ 8 mbp, often with large linear plasmids and particularly high GC content (Paradkar *et al.*, 2003). They also frequently contain giant linear plasmids carrying secondary metabolite gene clusters (Kinashi, 2011). These factors all reduce the genetic tractability of *Streptomyces* and make genome sequencing more complicated (Nakamura *et al.*, 2011; Gomez-Escribano, Alt and Bibb, 2016). Until very recently most sequencing chemistries would struggle with high GC organisms with large strings of G or C causing IonTorrent and Nanopore technologies to give greatly reduced yields. Currently, Illumina sequencing, PacBio and new Nanopore flow cells and software cope much more adeptly with these organisms and so the number of high quality, completed *Streptomyces* genomes is increasing rapidly ((Gomez-Escribano, Alt and Bibb, 2016)). With this there is an increase in the number of tools being developed to manipulate S*treptomyces* genomes such as pCRISPomyces (Cobb, Wang and Zhao, 2015; Kormanec *et al.*, 2019). Additionally, there is also more software to analyse *Streptomyces* genomic data sets. In genome mining alone there are many programs such as antiSMASH (Medema *et al.*, 2011), PRISM (Skinnider *et al.*, 2015), BAGEL (Van Heel *et al.*, 2018), CLUSEAN (Weber *et al.*, 2009) and RIPPER (Santos-Aberturas *et al.*, 2019) that can be used to obtain greater understanding of *Streptomycete* genomes, many of these were made specifically with this genus in mind.

1.3.2.1 DNA Replication in *Streptomyces*

It was not initially understood whether *Streptomyces* were a type of fungi, bacteria or something new entirely. Whilst we now know them to be bacteria, they do have several differences compared to a typical rod-shaped bacterium. They have twice as many genes as *E. coli*, linear chromosomes and have differences in their conjugation method compared to most prokaryotes including *E. coli*. Conjugation in *Streptomyces* is distinct from in *E. coli* where 30 *tra* genes are required while there is only one *tra* gene necessary in *Streptomyces* and there is also *spd* which is responsible for transfer of the plasmid through the mycelium, from where it was initially acquired, to more distant locations (Kieser *et al.*, 1982).The origin of replication (OriC) has a lower GC content than the rest of the chromosome and it maps to a similar location on the chromosome as *dnaA* and *gyrB* which encode DNA replication initiation proteins. The OriC contains 19 conserved DnaA boxes which allow interaction with DnaA proteins, which leads to the DNA unwinding and allowing for binding of the other DNA replication proteins (Pandza *et al.*, 1997; Messer *et al.*, 2001; Messer, 2002).

At the termini of *Streptomyces* chromosomes, similarly to telomeres in mammalian DNA, there are TIRs (terminal inverted repeats), these can range in size from a few base pairs to 1 mbp in *S. coelicolor* (Chan *et al.*, 2011). These regions are largely not conserved, and as they are repetitive, they can pose significant challenges to *de novo* genome assembly (Vembar *et al.*, 2016; Tørresen *et al.*, 2019).

1.3.2.2 Advances in genomics and the effect on *Streptomyces* biotechnology

One of the big drivers of biotechnology in recent years is the ability to target genomic changes to genomes and not have to rely on random mutagenesis or guided evolution of strains to produce the desired phenotypes. The rapid growth in sequencing technologies that has provided a vast wealth of sequences of both genes and whole-genomes that now populate databases has increased our knowledge of relating genotype to phenotype, for example StrepDB (http://strepdb.streptomyces.org.uk/). We now also have technologies such as the many variations on the CRISPR/Cas system that make targeted genetic modification in *Streptomyces* possible on a scale that it never was before. Due to reduced homologous recombination potential in *Streptomyces* species the techniques that are used for genome modification in other organisms are not feasible for *Streptomyces*. One of the major advances is using bacterial adaptive immune system components clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins to introduce targeted genome modifications (Hille *et al.*, 2018).There are a number of different classes of CRISPR/Cas systems and class II including CRISPR/Cas 9 have shown to be very useful in gene editing. Cas9 is a CRISPR guide RNA (crRNA)-dependent endonuclease with two nuclease domains, RuvC and HNH (Cho, Shin and Cho, 2018). RuvC is responsible for the cleavage of non-target DNA and HNH for cleavage of target DNA strands. Type II CRISPR-Cas loci contain a trans-activating crRNA(tracrRNA) that is required for pre-crRNA processing and target recognition (Wang, La Russa and Qi, 2016). Cas9 and crRNA form a complex that cleaves DNA at the target site (Ishino, Krupovic and Forterre, 2018). Multiple uses for CRISPR/Cas were found in the biotechnology sector. For eukaryotic cells CRISPR/Cas could make use of their non-homologous end joining mechanism to repair the double stranded breaks that it introduces. Whilst the system was less ground-breaking in bacteria than in mammalian biology, as there were already systems that performed a similar function, CRIPSR has still been widely used in bacteria. Unlike more genetically tractable bacterial species, *Streptomyces* was not able to be quickly and easily modified using homologous recombination methods so the prospect of CRISPR/Cas systems that work in *Streptomyces* was an exciting possibility for the *Streptomyces* community. The CRISPR/Cas based tool pCRISPomyces was developed to address a lot of the problems faced in modifying *Streptomyce*s genomes, it was shown to work in multiple *Streptomyces* strains deleting regions between 20 and 30 kb with an efficiency of 70~100% (Cobb, Wang and Zhao, 2015). It also has the advantage of being able to be used with spacers and editing templates made in a variety of methods including Golden Gate assembly and digestion/ligation.

1.4 Advances in Genome Sequencing

The history of DNA sequencing is relatively short to this point given how rapidly the advances in technology have come since the first attempts. Maxam and Gilbert's DNA sequencing method which relied on resolving the size of ^{32}P labelled DNA fragments to determine the sequence in 1977 was a huge advance in the field and led to numerous subsequent technological advances (Maxam and Gilbert, 1977).Since this technique many more, less

laborious sequencing methods came and went out of common use for large-scale whole-genome sequencing projects. The first of these was Sanger sequencing, which was developed shortly after, and used chain-terminating nucleotides with fluorescent dyes to determine the base incorporated at each position (Sanger and Coulson, 1975). Sanger sequencing is still used routinely for short <1kb sequencing projects due to its cost effectiveness and accuracy. Despite this, for whole-genome sequencing the Sanger method is not suitable as it is too time consuming and expensive on a large scale. For this, a method that allows for massively parallel sequencing reactions to occur was required, many different technologies satisfied this such as Roche 454 and Ion-Torrent but now Illumina dominates the field. These Next Generation (NGS) technologies have brought the price of sequencing per base down at a rate that massively exceeds Moore's Law as shown in figure 1.3. In this work Illumina sequencing has been used as it uses a reliable and accurate chemistry that produces short sequencing reads that are also paired, making the assembly process easier (Illumina, 2014). Illumina sequencing chemistry works in a "sequencing by synthesis" manner using clustering to generate paired reads. Isolated and purified genomic DNA is fragmented by transposases and sequencing adapters are added to the end of each fragment. Sequencing binding sites, indexes and complementary regions for binding to the flow cell are also added to each fragment by reduced cycle PCR. When introduced to the flow cell, which has millions of copies of two different oligos bound to its surface, the fragmented and adapter ligated DNA binds to one of the two oligo types. Each of the bound DNA fragments is amplified and then denatured to allow the loose end of the fragment to bind to the second type of oligo. Once again, this strand of DNA is amplified by a polymerase and denatured leading to two individual copies of the DNA molecule, each bound to the flow cell at opposite ends. This process is repeated until the flow cell is saturated and on the final round the denaturation is followed by the removal of the reverse strands. The first sequencing primer, in the adaptor region, is then extended by the addition of fluorescently tagged nucleotides which generates a base-specific signal whenever a new base is incorporated. After this first sequencing step has completed the fluorescent tag is washed away and indexing primer is used to generate the read specific index. After indexing of the first read the DNA fragment once again binds to the second oligo, and index two is read. The fragments are then amplified by PCR, denatured to from two single stranded fragments and then the reverse strand is sequenced as the first was. This process generates a large set of short reads that have defined spaces between forward and reverse strands, which allows for more robust assemblies due to the known pairing of reads (Dna, 2013; Illumina, 2014). Illumina sequencing chemistry is still improving regularly, and they have a variety of different systems that allows for both large sequencing centres to process vast quantities of genomes as well as smaller systems that are accessible to smaller academic laboratories. Illumina sequencing, whilst very accurate, has the disadvantage of being relatively short read length, ranging from 150-300 bp paired end reads, which is compensated for by the ability to obtain coverage of above 100X in many cases.

For some applications there is a requirement for very long read technologies, either to bridge gaps in a draft genome generated by short read technologies or indeed for *de novo* assembly. Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing is currently the gold standard in sequence technologies in terms of its ability to generate reads that lead to a completed genome when assembled. In the beginning of SMRT sequencing the read lengths were ~1.5 kbp and the error rate was ~13% which is relatively poor. The technology has improved greatly since then and now boasts average read lengths of ~15 kbp and whilst the

raw error rate of sequence reads remains the same this is now mitigated by the use of circular templates that can be sequenced multiple times. The downside of PacBio SMRT sequencing is that the technology is not particularly accessible to smaller laboratories with a PacBio RSII costing $700,000 USD in 2018 and PacBio sequencing provided by sequencing services it also considerably more expensive than other currently used methods. The other long read sequencing method is Oxford Nanopore Technologies (ONT), they have several devices currently available that use the same technology on different scales which accommodates different types of projects. Like SMRT sequencing ONT started off with shorter reads and low quality. As such this type of sequencing was not routinely used for *de novo* assembly of genomes until more recently when read lengths of over 2 Mb were regularly being reported by users and quality of >98% achieved regularly. In addition to these qualities both of these long-read sequencing technologies are in real time which allows data to be gathered as it is generated and used immediately which can be important especially in a clinical setting. Further advantages of the ONT Nanopore are that it is not sequencing by synthesis, so it actually reads the electrical signal given off by the DNA molecule as it is which allows for methylation to be detected.As it stands the three technologies that are favoured world-wide are Illumina, ONT and PacBio and these are often used in combination to resolve genomes.

All of these technologies have their place in various different applications, though it has been shown in many comprehensive whole-genome sequencing projects where genomes have been completed that often a combinatorial approach is required. This typically involves using both Illumina data alongside ONT data. Though any combination of long read and short accurate reads usually provides an improved assembly when compared to either technology alone (Beaton *et al.*, 2018; Goldstein *et al.*, 2019; McNaughton *et al.*, 2019).

1.4.1 Computational Challenges in Genomics

As has been previously mentioned there has been a rapid rise in the technology available to sequence genetic material as well as a huge downturn in the cost per megabase of sequencing (Prokop *et al.*, 2018). With whole-genome sequencing as well as many other uses for sequencing technologies becoming much more popular and commonplace the demand of for computational power and tools to analyse this data is also increasing. The general trend has been that the physical sequencing technologies have been evolving faster than their computational counterparts, this is well illustrated by recent developments in Nanopore sequencing. One of the main factors affecting the improved quality of the ONT sequencing reads was not a change to the chemistry of the system but rather a change to the software that reads and translates the output into base calls (Rang, Kloosterman and de Ridder, 2018). The idea that the science is held back by the computational capacity behind it is something that has followed the entire genomics era. With high throughput sequencing comes a greater need to have storage for large data sets, computational power to process the data into useful information and a population of computational biologists. Currently there are many permutations of programs that perform the main functions required to assemble a genome from sequencing reads and some expertise is required to choose the best methods for each project.

1.4.1.1 Base-calling

Base-calling is the step that takes the output from the sequencing technology, whether it be a series of images generated from an Illumina sequencer or a trace from an ONT device and turns it in to a series of letters. Illumina machines have a built-in base calling program and as standard it outputs data in FASTQ format. This means that the bases that have been called by the machine for each read are accompanied by the likelihood that this was carried out correctly in each instance. ONT provide a base calling program in their software, this software receives regular updates as it is a main contributing factor to the quality of ONT generated sequences. The first iteration of a ONT base caller was NanoNet which only survived one version and was generally poor as it gives lower sequence identity and less total output (Wick, Judd and Holt, 2019; Zeng *et al.*, 2020). Albacore took over from Nanonet and received several rounds of version updates before eventually being replaced by Guppy. Albacore and Guppy perform similarly when tested on identical data sets. The main difference between the two is that Albacore runs on CPU while Guppy can also run on GPU which makes it faster (Wick, Judd and Holt, 2019). These are the ONT-developed base callers that are packaged into MinKNOW, the software that ONT devices run on, for base calling as they are released but there are also some third-party base callers. The most notable of these is Chiron developed by (Teng *et al.*, 2018). Chiron uses deep learning and was trained on a relatively small data set to call bases from fast5 files, the main difference between this software and others is that it performs this on entire reads end to end without any event segmentation which is known to introduce errors. When various base callers were tested independently by Ryan Wick et al Guppy and Chiron came out on top in most aspects (Wick, Judd and Holt, 2019).

1.4.1.2 Read Error Correction

Read error correction is often not required for in-house generated Illumina reads due to their inherently high quality and high coverage of genomes. Outsourced Illumina reads can sometimes require some trimming to preserve only the highest quality of reads, though the level of error present here is more than compensated for by the coverage at any given error site. There are multiple trimming tools available for Illumina reads, in which various parameters can be selected to remove lower quality reads. A common method is to identify the base in the read at which the average read quality for the whole data sets drops below a chosen standard and remove everything beyond that point. This is simple enough to perform, though it has the downside of reducing the read length and as it is an average, some of the bases removed will have a high-quality score. Trimming using a sliding window approach is more specific, where bases will be removed from the read when the average score is below x over n consecutive bases. This reduces the likelihood that information will be discarded unnecessarily, as it is unlikely that a single base with low quality surrounded by high-quality calls will cause assembly issues in a medium-high coverage data set. A number of tools have been generated to perform these tasks in a GUI, notably via Galaxy servers (Goecks *et al.*, 2010; Droop, 2016). The need for read error correction is much greater with the long-read methods SMRT and ONT sequencing, as these methods are more error prone. The most comprehensive method of error correction in long read sequencing is to use a hybrid assembly method that combines long reads with Illumina sequencing to provide accuracy. This involves mapping the short reads onto the long and establishing a consensus for each base (Morisse, Lecroq and Lefebvre, 2020). The other option for long read error correction is self-correction, which is incorporated in many

long read specific assemblers including Canu (Koren *et al.*, 2017) and HGAP (Chin *et al.*, 2013). This method can be effective with a depth of coverage of 30x or more (Sedlazeck *et al.*, 2018) and is valuable especially with the intended field use of Nanopore devices. This works in the same way except the consensus is drawn based on overlapping long reads without the short-reads, assuming an error rate of ~10%, with a coverage of > 30x the likelihood of all reads having an error at the same locus are slim, as such the success in this method is dependent on the depth of coverage obtained. There are many tools to achieve this including PBcR-BLASR (Koren *et al.*, 2013), PBcR-MHAP (Berlin *et al.*, 2015) and MECAT (Xiao *et al.*, 2017).

1.4.1.3 Assembly

There are multiple genome assembly pipelines available using different algorithms, they all rely on having multiple sequence reads that overlap to produce longer sequences of DNA known as contigs this principle is shown in Figure 1.6

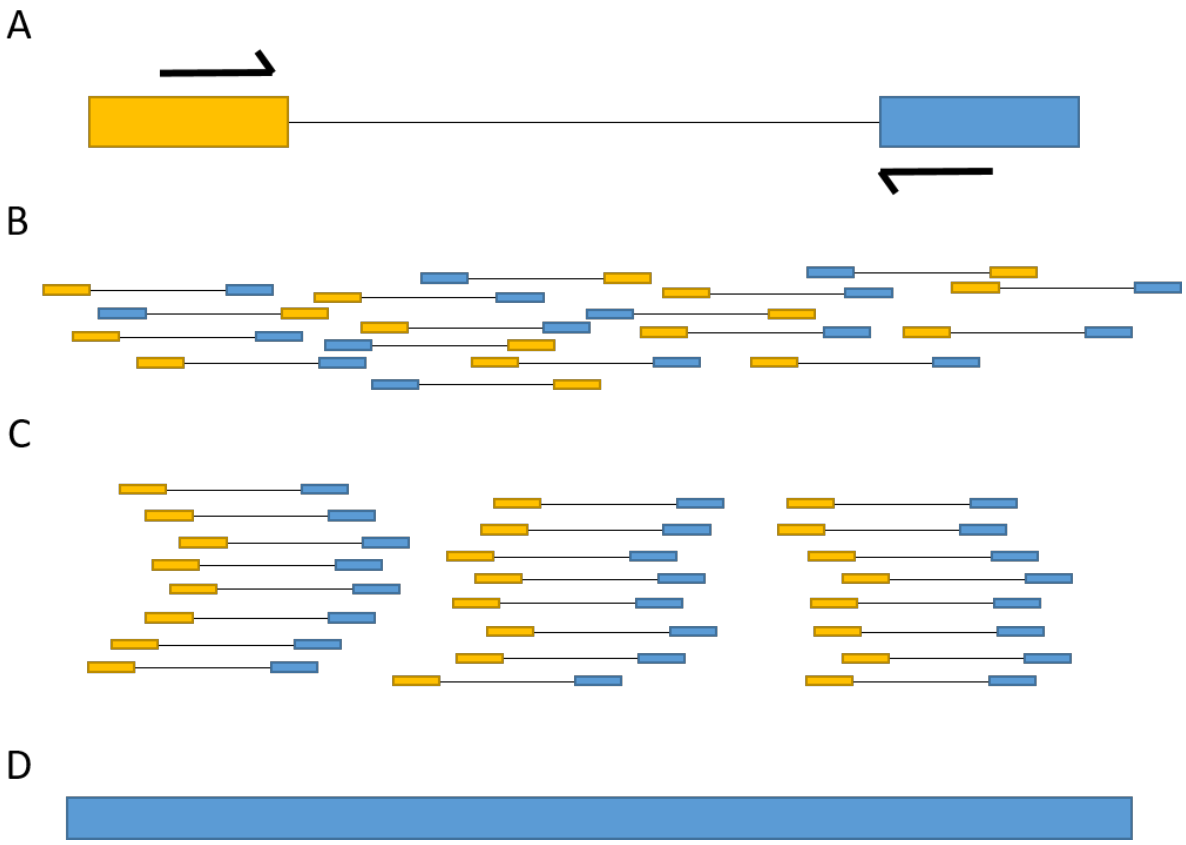


Figure 1.6. Assembly of paired end read data. A) shows one pair of reads with forward and reverse separated by a defined number of bases. B) A number of these paired end reads are produced by the sequencer and are given with per base quality score coded into the document C) the assembly algorithm attempts to find where single reads overlap by a defined number of bases D) A contig assembled from the consensus sequence of many overlapping short-reads.

1.4.1.3.1 SPAdes

Many assemblers, including some of those used in this work employ a De Bruijn graph based approach to *de novo* assembly (Compeau, Pevzner and Tesler, 2011; Langmead, 2014). Mathematical graph theory is the use of structures to examine the relationship between objects. Many assemblers, including SPAdes, use some form of a graph-based approach (often based on de Bruijn graphs) in some form. Figure 1.7 gives examples of graph construction from a simple "genome." SPAdes uses a multisized de Bruijn graph which is an advance on the simple de Bruijn graph used in Velvet that removed "bubbles" in assemblies. SPAdes generates multiple de Bruijn graphs using multiple different k-mers and uses the results from multiple graphs to generate a single, improved, graph that then undergoes further assembly improvements (Bankevich *et al.*, 2012). After graph construction SPAdes the performs k-bimer adjustment which accurately estimates the distance between k-mers in the genome which allows for effective assembly of paired reads, then a paired assembly graph is constructed and then the contigs are constructed.

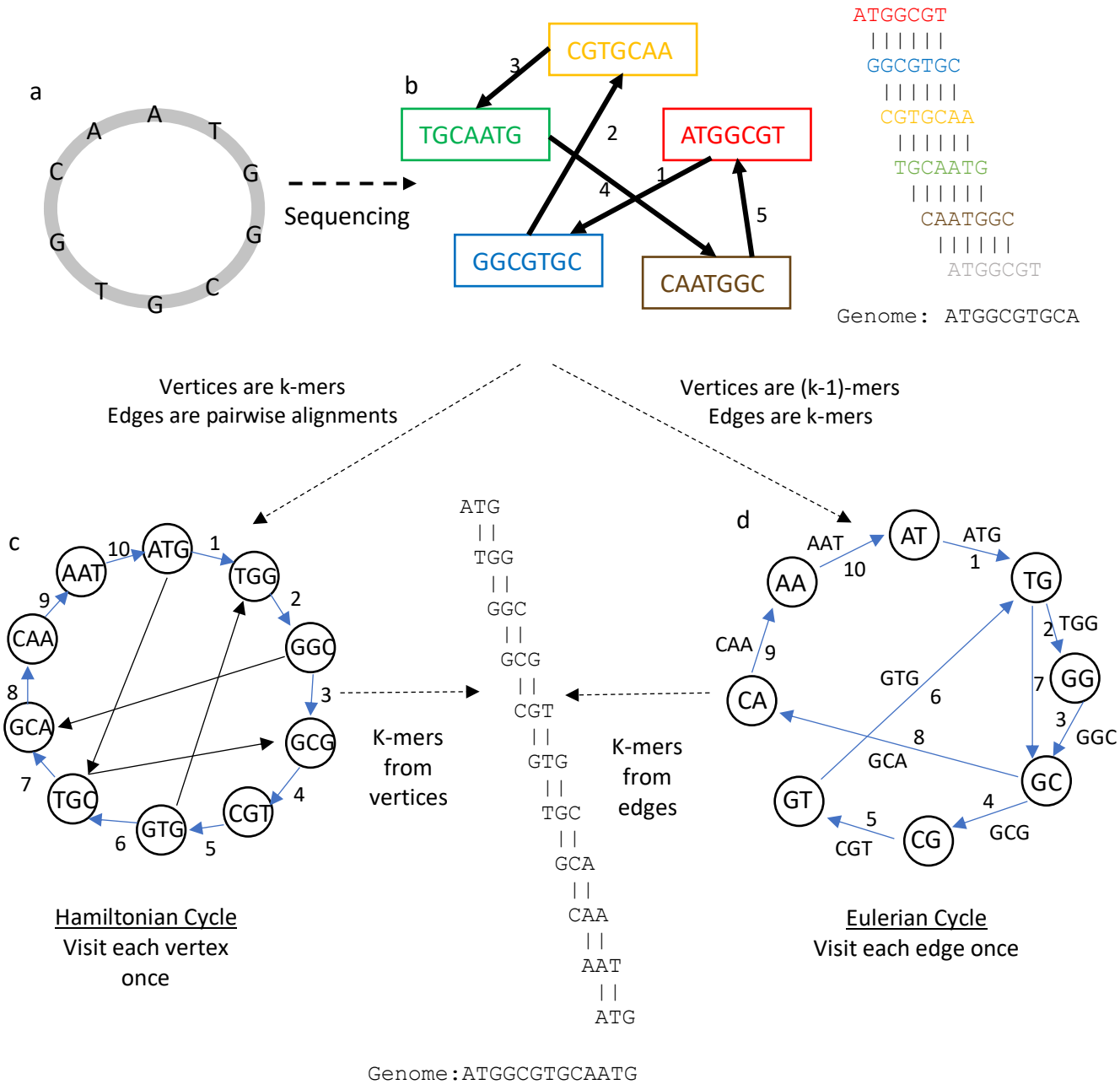


Figure 1.7. Examples of de Bruijn graphs. Showing two of the types of de Bruijn graphs, Hamiltonian and Eulerian. a) a circular genome. b) showing reads represented as nodes and alignments as edges. "Walking" along the edges in order generates the circular genome. c) reads have been split into k-mers The blue edges represent the Hamiltonian cycle in which each k-mer (from nodes) is shifted by one position. d) each k-mer prefix and suffix is a node and each k-mer is a edge.

1.4.1.3.2 Long Read Assemblers

Whilst SPAdes can accommodate input from long read technologies, via Unicycler (Wick *et al.*, 2017) there are several assembly tools that have been developed with long, error-prone reads in mind. One of which is Flye (Kolmogorov *et al.*, 2019), this works in a similar way to SPAdes in that it is also a graph-based assembly but this tool uses repeat graphs rather than de Bruijn graphs. One of the benefits of long-read sequencing is to resolve repetitive regions in genomes that short-read sequencing cannot, however, many assemblers still struggle to assemble these regions even with long reads (Du and Liang, 2019). Flye was developed to improve assembles in this aspect, and so it modifies the way it constructs the assembly graph compared to a De Bruijn graph (Kolmogorov *et al.*, 2019). The general concept of Flye assembly graphs is represented in figure 1.8. Instead of assuming a k-mer represents the same originating sequence that has been sequenced multiple times and compressing those reads down, Flye takes these kinds of reads and forms them into disjointigs, as shown in figure 1.8c. The disjointigs are then concatenated (figure 1.8d), probably incorrectly and formed into a repeat plot as shown in figure 8f. The repeat plot is constructed by representing repeats on the main diagonal and points where these coincide are used to infer convergence points on the repeat graph (Figure 1.8g). Flye then maps the reads onto the repeat graph 8h. Flye deals with bridged and unbridged repeats in different ways. Firstly, to explain the difference, a repeat is said to be bridged if there is at least one read that completely covers one instance of the repeat and unbridged if such a read does not exist (Kamath *et al.*, 2017; Kolmogorov *et al.*, 2019). In a comparison of long read assemblers Canu and Flye outperformed others but were relatively similar to each other in most cases. Flye slightly outperformed Canu in terms of genome contiguity and run time, while Canu uses less memory (Wick and Holt, 2019). It remains that there will be no one pipeline that could be considered the best of all sequencing projects and in any case more than one method would need to be used to ensure and compared on a project-by-project basis to ensure the best assembly.

Other types of assemblers use linguistic algorithms to resolve the patterns in the genome. Canu (Koren *et al.*, 2012), which was published in 2017 is a long-read assembler specifically designed to deal with the error prone data from Nanopore sequencing. Canu combines Celera Assembler (Myers, 2005; Miller, Koren and Sutton, 2010), MinHash Alignment Process (MHAP) (Berlin *et al.*, 2015) and PBcR (Koren *et al.*, 2013) to provide a *de novo* assembler with reduced requirements for coverage with Oxford Nanopore and PacBio data that can give better results separating repeat regions and haplotypes. MinHash is a mechanism to estimate the Jaccard Similarity between two individual data sets by mapping them onto one another (Ondov *et al.*, 2016).This algorithm which was originally developed to determine the similarity in two web pages allows for large pieces of data to be reduced to a “sketch” which makes comparisons more efficient (Broder, 2000). This method has potential to collapse repeat regions, assuming that each iteration of the repeat is a single region sequenced multiple times. To overcome this Canu recomputes the overlaps after unsupported overlaps are removed. The first iteration of this process detects probable errors in the reads and then the alignment scores are recomputed with the temporarily ‘corrected’ reads. In this way when the graph is constructed there is a higher proportion of very similar haplotypes and repeats being resolved (Koren *et al.*, 2013).

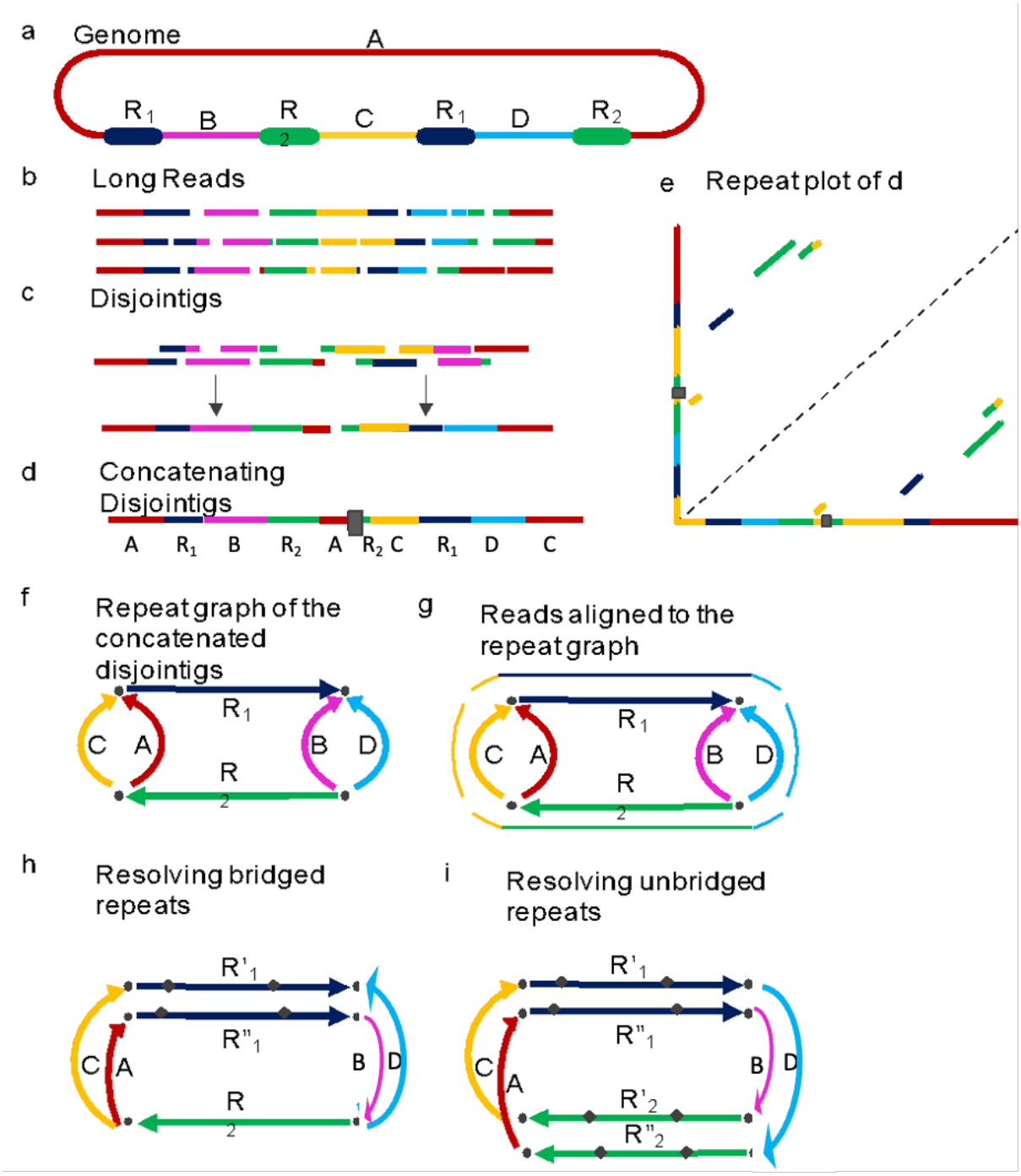


Figure 1.8. Flye Assembly Method. A) A genome with two sets almost identical repeat regions represented by R_1 and R_2. A, B, C and D are non-repetitive regions. B) Set of reads from the genome in a. C) A representation of how the reads from b may be assembled (incorrectly) D) concatenated disjointigs from C. E) Repeat plot of the concatenate shown in d. with repeats represented as diagonal alignment paths. F) Repeat graph where the clustered endpoints generated from coinciding projections in the repeat plot shown in e. G) reads aligned to the repeat graph. H) resolving the bridged repeat R1, with the small differences in the two copies represented as R'_1 and R''_1. I) Graph resolving unbridged repeat R_2 with the two copies represented as R'_2 and R''_2. Adapted from figure 1 in Kolmogorov, M., Yuan, J., Lin, Y. *et al.* Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37,** 540–546 (2019).

1.4.1.4 Annotation

Annotation tools are essential for identifying the different regions of a genome and assigning a function to both coding and non-coding sequences (Overbeek *et al.*, 2014).There are several commonly used genome annotation tools including RAST (Aziz et al. 2008), PROKKA (Seemann, 2014) and several genus-specific pipelines.

1.4.1.4.1 RAST

RAST or Rapid Annotations using Subsystems Technology is a microbial annotation server that is based on SEED. SEED is a platform that integrates genomic data from multiple sources including literature data, metabolic modelling data, user annotations and public genomes annotated by RAST. RAST has a multistep process to identify and annotate genes, GLIMMER3 which is based on interpolated Markov models to pick out protein-coding sequences is used to make initial identifications of genes that are used to identify strains that are phylogenetically similar to the genome being annotated (Salzberg *et al.*, 1998; AL Delcher S Kasif, O White and SL Salzberg, 1999; Delcher *et al.*, 2007). Genes identified by this means are then compared to known k-mer sequences from proteins that have been designated into a subsystem. Gene candidates that can be assigned a functional role based on these k-mers are labelled as protein encoding genes (PEG). This process is repeated iteratively until no new gene candidates are found. Candidates that are not designated PEGs from this process are assigned functions based on BLASTP similarities if possible. Furthermore, candidates that fit neither of these criteria but also do not overlap another called gene are included, which should ensure novel genes are not excluded. Selenoproteins and pyrrolysoproteins are identified using a custom algorithm as BLAST has been found not to be sensitive enough to differentiate between these and similar genes. tRNA and rRNA genes are found using 'search_for_rnas' from Neild Larsen and BLASTN with a set of RNA databases. This all comes together to generate a comprehensive annotation that can be explored and compared using the SEED viewer (Aziz et al. 2008; Overbeek et al. 2014). RAST is not designed for the annotation of eukaryotic genomes.

1.4.1.4.2 PROKKA

PROKKA is a pipeline that was largely developed to fill a gap in the market for an annotator that is easily integrated into a genomics pipeline as many others including RAST work on a web-based server and thus cannot be inserted into a custom pipeline by the user. PROKKA identifies coding regions and then compares them to known databases to assign a function. PROKKA uses Prodigal (Hyatt *et al.*, 2010) to identify protein-coding regions. It then compares these gene candidate regions to known databases starting at the smallest, most specific, and trusted, database and working up and out in a hierarchical manner. Hidden Markov model based protein databases are searched using the program HMMER 3.1 (Eddy, 1996). PROKKA has the option for the user to specify —usegenus which will use a genus-specific database for annotation. This is useful if you are confident in the assignment of the genome you sequenced. Where no similarity is found the coding region is designated as a hypothetical protein. Similarly, to RAST, PROKKA is designed to annotate bacteria, archaea and viruses not eukaryotic

genomes. Eukaryotic genomes cannot be annotated using PROKKA as it only provides prokaryotic databases from which to identify sequences.

1.4.1.4.3 FunGAP

There are a wide range of eukaryotic genome annotation tools available especially for human and mammalian genome analysis (Schweikert *et al.*, 2009; Hoff *et al.*, 2016). There are also several fungal or yeast specific pipelines that are of more interest here since they should give more accurate annotations by using specific databases, in the same way RAST and PROKKA do with bacterial domains. FunGAP (Min, Grigoriev and Choi, 2017) is one of the fungal specific annotation pipelines that exists, it also works in the same general manner of identifying protein-coding regions and assigning them a function based on similarity though it is more complex than the other software mentioned in both its algorithm and use. From a user's point of view, FunGAP is less friendly as it doesn't have a graphical user interface (GUI) as RAST does and it also requires more input. Other annotation tools mentioned so far take a FASTA file with the assembled genome and perhaps some metadata data as optional extras. FunGAP requires the FASTA assembly file, Illumina mRNA reads in FASTQ format and a protein database in FASTA format. In some cases, these input requirements could be prohibitive, if you have not performed Illumina mRNA sequencing along with your genome sequencing. In some cases, it will be possible to use the mRNA data from a similar strain to facilitate annotation by FunGAP where you don't possess this data for your own sample, but this may lead to inaccuracies. FunGAP uses three programs for gene prediction Augustus (Stanke *et al.*, 2006), Braker (Hoff *et al.*, 2016) and Maker (Cantarel *et al.*, 2008). FunGAP only uses non-overlapping genes by evaluating all candidates that do overlap and choosing only the one that has the most similarity to another known protein. All predicted genes are then given an evidence score by combining the scores from BLAST against the provided database, Pfam and BUSCO. FunGAP was analysed using multiple fungal genomes and determined to provide better annotation using the combination of predictors than any one of them alone (Min, Grigoriev and Choi, 2017).

1.4.1.4.4 YGAP

Another specific, eukaryotic annotation software package is the yeast genome annotation pipeline (YGAP) (Proux-Wéra *et al.*, 2012). Which is specific to yeast genomes as FunGAP is to fungal genomes, but this pipeline is more user friendly in that it works in a more similar way to RAST with a web-based GUI and required the input of only the assembly scaffold in fasta format. There is also the option include the reads file and the contigs file if desired. YGAP uses the yeast gene order browser (YGOB) database to search and assign function (Byrne and Wolfe, 2005). YGOB organises its data into a matrix consisting of tracks and pillars. For pre-WGD species there is a single tract and two tracks for post WGD strains. The pillars consist of sets of homologous genes. The first step in the YGAP pipeline is using tRNAscan-SE (Lowe and Eddy, 1996) to detect and annotate tRNA genes. The tRNAs detected at this point will not be allowed to overlap with any other gene that is a predicted protein-coding gene later in the process. TBLASTN searches are used where proteins of each pillar in the YGOB is used as a query to assign predicted functions to the protein-coding genes.In pre-WGD strains only one locus can be assigned to an existing pillar but in post-WGD strains two loci can be assigned to a single

pillar. One of the reasons given earlier for bacterial annotators not being effective with yeast is the genetic structure, as such YGAP has a method for annotating introns. Since introns are rare in yeasts they are only looked for where another gene in the pillar that gene has been assigned to also contains an intron. For comparison, a S288c annotation performed using YGAP was compared to the annotation of the same strain carried out by Augustus, as it has been trained on *S. cerevisiae* data sets. This comparison found that 181 more genes were correctly predicted with YGAP than Augustus, there were also fewer false positives and negatives from YGAP, both methods had a similarly high rate of false negatives for intron prediction.

1.4.1.5 Assembly Quality Assessment

There is no real way to definitively decipher how correct a *de novo* genome assembly is without wet lab techniques however, there are several ways in which to assess the quality of an assembly computationally. QUAST is a program that compiles numerous basic metrics from the final assembly fasta file to determine the contiguity of a particular assembly (Gurevich *et al.*, 2013). The most notable of the metrics that QUAST outputs is the total genome size, the number of contigs that the assembly is comprised of and the N50 value. N50 is the number of bases that 50% of the total number of bases in the assembly are housed in a contig of that size or larger (Mäkinen, Salmela and Ylinen, 2012). Whilst these measurements can't prove a genome has been accurately assembled, they can give some indication as to what methods may be most appropriate. For example, a genome assembly with 2000 contigs is probably worse than one with three, though not necessarily. Another script for assessing genome assembly completeness is BUSCO (Simão *et al.*, 2015). BUSCO is designed to complement other quality metrics by estimating what proportion of genes are found, that are expected to be present based on their "single copy" representation in the taxon of interest the assembly has captured. This method is quite robust as we know that a genome in a single contig may not necessarily be correctly assembled but if it has all of the expected essential genes then at least most of the metabolic potential should have been captured. For many projects, this will be the most important outcome, as such it is a valuable tool. The downside to this, is that the accuracy of this is dependent on the knowledge of the specific organism that is being assessed with the potential of novel and phylogenetically distinct organisms being difficult to assess in this way. Another method of assembly quality assessment developed specifically for assemblies made using long read sequencing, such as Nanopore and PacBio is Ideel. Ideel determines the frequency of proteins in the assembly that are truncated or contain InDels (insertion/deletion events) as an artefact of the, typically more error prone, long read sequencing chemistries. This analysis compares the predicted protein coding regions in an assembly to a provided database of proteins and determines the frequency of proteins that appear to have been truncated or extended. Ideel does not differentiate between proteins that have been truncated as an effect of poor assembly, or if there has been a genuine truncation in the genome. It will however indicate quality differences between versions of the same assembly, for example corrected and uncorrected, and more importantly gives an indication as to whether further interrogation of truncated genes is required. None of these metrics can ensure a high-quality genome but combined, when used in the correct way can indicate the correctness of an assembly in that with "good" results from both QUAST, Ideel and BUSCO one could predict with a high likelihood that a genome is both contiguous and contains all or most of the expected coding regions.

1.4.2 Database and Bioinformatics Pipeline Availability

Importantly, with the vast improvement in genome sequencing and the increase in genetic sequence data available there has also been an increase in the number of tools available to analyse these data. This includes software to analyse sequencing data quality, assembly and mapping algorithms as well as a variety of tools to assign function and meaning to genomic data. A large contributor to the ease of use of a lot of bioinformatics analysis software is Galaxy, the several instances of which allow for GUI based analysis to be run simply from any computer (Goecks *et al.*, 2010). Another resource to avoid users requiring significant computational power of their own or knowledge of command line is CLIMB (the Cloud Infrastructure for Microbial Bioinformatics) (Connor *et al.*, 2016)

1.4.2.1 Read Mapping and Variant Detection

Mapping of reads to a reference can be useful in many circumstances, for example when reads have been obtained for a variant of a strain to which there is already a complete genome assembled this can help with assembly and also in the detection of short nucleotide polymorphisms (SNPs) and insertions and deletions (Indels). This technique can also be used to identify regions of the genome that may be duplicated or to manually resolve errors in assembly.
The Subread aligner (Liao, Smyth and Shi, 2013) uses seeds (a short part of a read) to match to a part of the genome. This is because to align the entirety of each read to the genome would be computationally very expensive. This algorithm is fast and accurate because no individual is required to match the genome directly, but the final location will only be assigned if multiple seeds agree. This is termed the Seed-and-vote strategy. It works for both long and short-reads, with short-reads (<160 bp) overlapping subreads are used in the algorithm. This strategy attempts to improve upon the more common method of read mapping which also involves seed aligning but since seeds usually align to more than one location on the genome, the rest of the read then needs to be used to determine which of the multiple possibilities each read matches, this is the most time-consuming stage, and it increases with read length. In addition, this method can be complicated by read errors, indels and in the case of RNA-seq, exon junctions.

One of the most commonly used aligners is Bowtie2 (Langmead and Salzberg, 2013) which uses the full-text minute index based alignment method from Bowtie and then complements this with dynamic programming to allow for gapped extension stage without too much of a time penalty (Ziemann, 2016). The dynamic programming allows for the alignment program to be broken into smaller problems recursively and each step in a route matrix of the alignment can be given a score and the information used to generate that score can be discarded before the next decision is made and furthermore since only the score of the previous step is used to calculate the next the score from point A can be discarded after calculating point B and so on. In this way sublinear space can be achieved.

1.4.2.2 Genome viewers

Of vital importance to understanding sequenced genomes is the ability to view not only the assembled contigs but also the raw reads and how these may map to the assembly. Despite the many advances in genome sequencing and analysis described here, there is still some level of requirement for manual curation in these pipelines. Various aspects of genome assembly pipelines can be manually "quality controlled' with proper utilisation of a genome viewer that can allow for SNPs, repeats, annotations and alignments (Pavlopoulos *et al.*, 2015). There are several tools that allow viewing of assembled genomes, the most basic of which being text editors that simply show the genetic code that comprises their genome. Software such as Artemis and Tablet are designed for the viewing of genome annotations and assemblies respectively and therefore have several intrinsic features that aid in the curation of genome assembles.
Artemis is a popular genome viewer that is a freely available Java software from The Sanger Centre (Rutherford *et al.*, 2000). It allows the import of data in multiple formats to support the visualisation of both assemblies and annotated genomes. It also allows for all six reading frames to be visualised which can give important context to bacterial operons. In addition to this it allows for the mapping of raw read data onto the assembled genome which gives a visual representation of coverage and the ability to visualise errors in the reads and/or the assembly. Tablet, another viewer from the James Hutton Institute, is similar in many ways to Artemis (Milne *et al.*, 2016). The main advantage of Tablet over Artemis as it pertains to this work is the improved visibility of mapped reads with zoom and colour toggles highlighting particular features such as SNPs so that they can be much more easily identified.

Integrated Genome Viewer (IGV) and Integrated Genome Browser (IGB) (Robinson *et al.*, 2011, 2017; Thorvaldsdóttir, Robinson and Mesirov, 2013) from the Broad Institute offer the ability to browse genomes much the same as the previously mentioned tools. As with Artemis and Tablet IGB allows for RNA-seq data to be visualised in the context of the reference genome assembly, and IGB can produce coverage graphs and data for both this and for NGS reads.

1.4.2.3 Secondary metabolite cluster prediction

One of the main reasons that such a variety of *Streptomyces* strains have their whole-genome sequenced is that there are many groups interested in finding secondary metabolite gene clusters that may or may not be known of by phenotype alone in a laboratory setting. As such the tool antiSMASH which searches complete or draft genomes to find regions that have sequence similarity to biosynthetic gene clusters (BGCs) is very helpful. Antibiotics and secondary metabolite analysis shell (antiSMASH) was developed to identify and analyse BGCs in both bacterial and fungal genome sequences. Now on version five antiSMASH can identify several different cluster types including polyketides, non-ribosomal peptides, bacteriocins, fungal RiPPs, lipolanthines, oligosaccharide antibiotics etc (Medema *et al.*, 2011; Blin *et al.*, 2017; Blin, Shaw, *et al.*, 2019) AntiSMASH is complemented by antiSMASH database (antiSMASH-DB) which houses all of the antiSMASH predictions for all genomes published on NCBI before May 2016, this allows for rapid cross matching of results and searching for particular clusters across a number of species (Blin, Pascal Andreu, *et al.*, 2019). The more recent iterations of antiSMASH are also able to detect biosynthetic gene clusters from fungal

genomes with a comparable level of success compared to the next-best predictor for fungi, SMURF (Khaldi *et al.*, 2010)

1.4.2.4 Whole-Genome Phylogeny Tools

Since sequencing first was possible phylogeny has been estimated by 16S for bacteria and 18S for fungi, despite the numerous flaws in this method (Case *et al.*, 2007; Větrovský and Baldrian, 2013; Zhulin, 2016). The use of small ribosomal subunit sequences to classify bacteria, archaea and fungi was a ground-breaking discovery that reshaped how we think about microbial taxonomy and necessitated re-naming and re-classification of many strains. Whilst this is invaluable for classifying microorganism in a standardised way it often fails to give any real insight into the strain that has been identified (Clarridge, 2004). Due to the high proportion of important genes that can be encoded on plasmids or other mobile genetic elements and transferred between strains, species and even domains knowing the 16S sequence often is not useful information. Furthermore, due to the high-level of 16S sequence conservation the resolution is low and usually only able to distinguish at genus level. In addition to these downfalls there is also the fact that it is becoming increasingly cheap to obtain whole genome sequences for a microorganism, which puts more informative methods such as whole-genome or multi-locus-based phylogeny within reach . One popular tool for this, particularly in actinomycetes, is AutoMLST which is an online tool with a graphical user interface (GUI) that takes draft or complete genomes, identifies a particular set of genes and uses these to make a phylogenetic tree (Alanjary, Steinke and Ziemert, 2019).

1.4.2.5 Phenotypic Analysis

The publication of Bergey's Manual of Determinative Bacteriology in 1923 based microbial taxonomy on observable phenotypes, and phenotypic classification is still a major component of taxonomic classification. This is because how an organism behaves in observable and measurable ways was the only mechanism of identification pre-genomics and still is of great importance, particularly to industry. The use of sequencing and genomic analyse to predict phenotype is an increasingly powerful mechanism, it does however still require confirmation by wet lab experiments. One of the downsides to using traditional phenotyping wet lab experiments to understand the capabilities of an organism is how time consuming it is to obtain information on the full spectrum of biochemical capabilities. In addition to this growth of flocculating or pelleting organisms can pose difficulties during the acquisition of growth data as sampling bias can render measurements inaccurate and optical density (OD) if often not a viable method of measurement.

1.5 Industrial Strain Phenotyping

Genomic information can be utilised to expedite the phenotyping process and direct what tests are carried out, but high through-put phenotyping can also be used in conjunction to allow for large-scale screening that can attempt to match the volume of potential targets revealed by whole genome sequencing.

1.5.1. Biolog Phenotypic Arrays

Biolog phenotypic microarrays have been developed that make it possible to assay the growth of an organism in multiple conditions in parallel. The Biolog system provides 96 well plate format growth assays that test a strains ability to grow on different carbon sources, nitrogen sources, phosphorus and sulphur, osmotic stress, pH effects and chemical sensitivity (Mackie *et al.*, 2014). The desired condition is lyophilised in each well allowing the strain to be tested to be suspended in a custom solution providing the strain's other growth requirements as well as a tetrazolium-based dye.The dye is reduced during respiration which catalyses a colour change which is in turn measured by the Biolog at a designated time interval. In this manner ~5000 strain and nutrient condition combinations can be measured simultaneously in a standardised assay (Biolog, 2007). Growth data is therefore observed via the proxy of respiration, as such this method does not necessarily tell us under which conditions most cells made in, but it does show how much the total culture respires which has been shown to be an adequate proxy for overall growth (Greetham, 2014). Due to the nature of information captured being a camera measuring the colour intensity of the dye in the whole well, issues faced with OD in organisms that are prone to forming an aggregate are irrelevant. The Biolog system can be utilised in multiple ways, including microbial identification, identification in polymicrobial communities (Smalla *et al.*, 1998) and in the identification of suppressor mutants (Bochner, 2009).

1.5.2 Relating Phenotype to Genotype

Having the whole genome sequence of an organism as well as a vast amount of data on its phenotype presents the opportunity to further understand the organism by relating these two pieces of information together to be able to understand how genomic differences are affecting the organism. This is typically done via a genome wide association (GWAS) study.

1.5.2.1 Genome Wide Association Studies

Genome wide association studies (GWAS), are often used in human genetics to identify disease-causing genes, look at the genetics of an organism including SNPs and find correlations between genomic variations and trait variations by comparison to other samples (Tam *et al.*, 2019; Tsunoda, Tanaka and Nakamura, 2019). The simplest example would be if two strains were identical phenotypically apart from the ability of one to grow without a given amino acid and the same two strains also being genetically identical apart from one SNP. A GWAS would lead to the hypothesis that this SNP might be responsible for the metabolic difference. Multiple platforms for human GWAS (hGWAS) exist including PLINK (Purcell *et al.*, 2007) and FaST-LMM (Lippert *et al.*, 2011) but it is more complicated with microbial genomes because of the much higher levels of genetic plasticity. Microbial genome-wide association studies (mGWAS) are still relatively new and as such there is not currently a particular, user friendly complete pipeline for this time of analysis which potentially reduces its use (San *et al.*, 2020). There are three main types of variation that are possible between microbial genomes, and they are SNPs, gene presence/absence and copy number variations. There are also multiple types of phenotypic variation, including improved or reduced growth in a given condition, increased or decreased pathogenicity, loss of gain of a particular function such as sporulation and many more (Power, Parkhill and De Oliveira, 2016). In addition, whilst many phenotypes may already have well characterised genotypes associated with them many have not been elucidated fully or at all

meaning that binary result is impossible in most cases (Dutilh *et al.*, 2013). As such, traits that give the most obvious and distinct phenotypic changes are the easiest to attach to a genomic locus difference, furthermore GWAS is easier on larger data sets from which more statistical significance can be obtained (Lees *et al.*, 2016)

There are a number of mGWAS tools available each working in different ways with different implementations. Sequence Element Enrichment (SEER) and pySEER (Lees *et al.*, 2018) are popular and available for free in GitHub. SEER can take input data as either a *de novo* assembly or raw sequencing reads and works on tens of thousands of genomes. SEER and pySEER look for multiple types of variation including InDels and deletions. The algorithm used looks to avoid the need to create a core alignment which can lead to inaccuracies whilst also being scalable to use on big data sets. They were trained on data from *Streptococcus pneumoniae* and *Streptococcus pyogenes* and was found to be able to identify regions pertaining to antimicrobial resistance with a high level of accuracy (Lees *et al.*, 2018)

1.6 Aims of the present work

The main aims of this work are two-fold, firstly to genetically characterise and phenotypically assess candidate strains of brewer's yeast *Saccharomyces cerevisiae* to determine the traits that confer most productivity in the bio-process specific to our collaborators where corn-mash is the feedstock in a bio-ethanol fermentation. This should assess the success of previous efforts made to expand the carbon utilisation profile of the *S. cerevisiae* strains to best exploit a variable and unpredictable feedstock. Secondly to analyse a lineage of industrial *Streptomyces* strains to identify their improved ability to produce a desired compound. Four strains of *Streptomyces lavendulae*, each known to differ in their ability to produce the desired end product, were sourced from public culture collections. Testing in the facilities of the industrial collaborators Ingenza did not produce the anticipated results, so further interrogation of the strains was required to inform future modification to provide improved end production yield

In both cases, in-house whole genome sequencing including both short-read Illumina chemistry and long read Nanopore sequencing to produce high quality genomes that can be used to assess genetic differences, including SNPs, indels and rearrangements between strains in the lineage and relate these to the observed phenotypic differences.

Chapter 2 Materials and Methods

2.1 Strain Acquisition

2.1.1 *S. cerevisiae*

S. cerevisiae strains, Ethanol Red, Angel, IGZ008, IGZ093, IGZ094, IGZ0132, IGZ0133, IGZ0134 and IGZ0135 were obtained from public repositories or Ingenza Ltd.

2.1.2 Streptomyces

Four *Streptomyces* strains were obtained from the following sources *Streptomyces ardus* (DSMZ 40603), *Streptomyces lavendulae* (S. caespitosus) NRRL 2564, *Streptomyces caespitosus* AN-9 (NRRL 12513), *Streptomyces caespitosus* T-17-135 (NRRL 12508). This group of strains is reported to have varying abilities to produce members of the mitomycin family, which is pertinent to the compound of interest.

2.2 Culture of Microorganisms

All media was prepared as detailed below and sterilised by autoclaving at 121^{O}C for 15 minutes and was then stored at room temperature until use. In the case of agar, medium was placed in a water bath for 1 hour until it reached 55^{O}C for the addition of antibiotics and/or pouring.

2.2.1 Culture Media

YPD medium was prepared as follows peptone 2% (w/v) (Fisher Scientific BP1420-2), 1% (w/v) yeast extract (Sigma Y1625), dextrose 2% (w/v) (Fisher Scientific G/0500/60) and 2% (w/v) agar (Fisher Scientific A/1080/53) for solid medium (Harbor, 2010).

LB medium was prepared as described 1% (w/v) Tryptone (Melford T60060), 0.5% (w/v) yeast extract, 1% NaCl (Fisher Scientific S/312/60) and 2% (w/v) agar for solid medium (Son and Taylor, 2012).

TSB medium was prepared as follows 0.3% (w/v) Soy (Holland & Barrett 29185/04), 1.7% (w/v) Tryptone, 0.5% (w/v) K_2HPO_4 (AnalaR 103494G), 0.5% (w/v) NaCl, 0.5% (w/v) dextrose and 2% (w/v) agar for solid medium (Sa-uth *et al.*, 2018).

SFM media was prepared as follows, UK 0.3% (w/v) soy, 1.7% (w/v) tryptone, 0.5% (w/v) K_2HPO_4 0.5% (w/v), 0.5% (w/v) NaCl, 0.25% (w/v) dextrose and 2% (w/v) agar (Aínsa, 2000).

SC Medium was made using Yeast Nitrogen Base (Sigma Y0626) according to manufacturer's instructions

D- Glucose for Biolog (24x) was prepared as described by Biolog manufacturer protocols with 43.248 % (w/v) dextrose.

2.2.2 Antibiotic Concentrations

Antibiotic	Concentration
Zeocin (InvioGen ant-zn)	300 µg/mL
G418 (Life Technologies 10131-019)	200 µg/mL
Ampicillin (Melford A40040)	100 µg/mL

2.1.3 Buffers and Reagents

TAE Buffer
50X stock solution:
50 M ethylenediaminetetraacetic acid (EDTA) disodium salt (Melford E57020) pH 8.0,
2M Tris-base and 1 M glacial acetic acid.

2.3 Plasmids and Primers

2.3.1 Primers

Name	Sequence
KanR_2_FWD	TGGGTAAGGAAAAGACTCACG
KanR_2_REV	TCATCGAGCATCAAATGAAAC

2.3.2 Plasmids

Name	Size	Resistance Genes	Plasmid Map
pING076	10821	AmpR, KanR	Figure 2.1
pING077	10642	AmpR, KanR	Figure 2.2
pING074	8138	AmpR, BleR	Figure 2.3

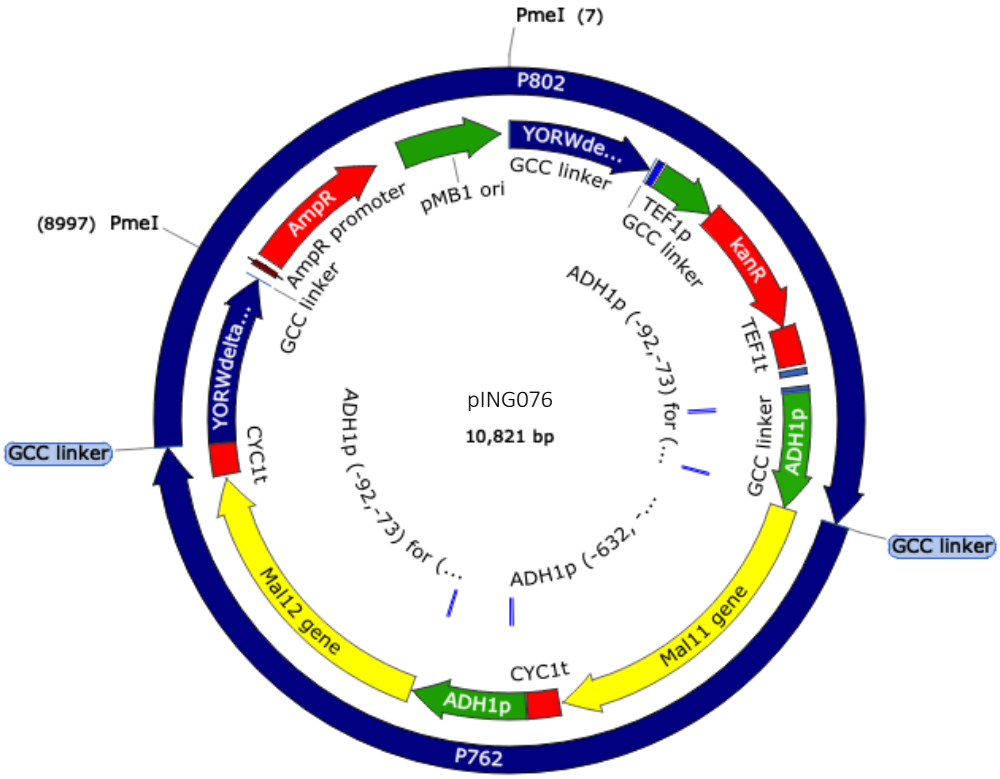


Figure 2.1. Plasmid map of pING076. This plasmid contains maltose transport and utilisation genes with ampicillin resistance gene for maintenance in *E. coli* and kanamycin resistance for selection in *S. cerevisiae*. Contains MAL11 and MAL12 for the transport and utilisation of maltose under the expression of the constitutive ADH1 promoter, flanking sites homologous to the *S. cerevisiae* genome.

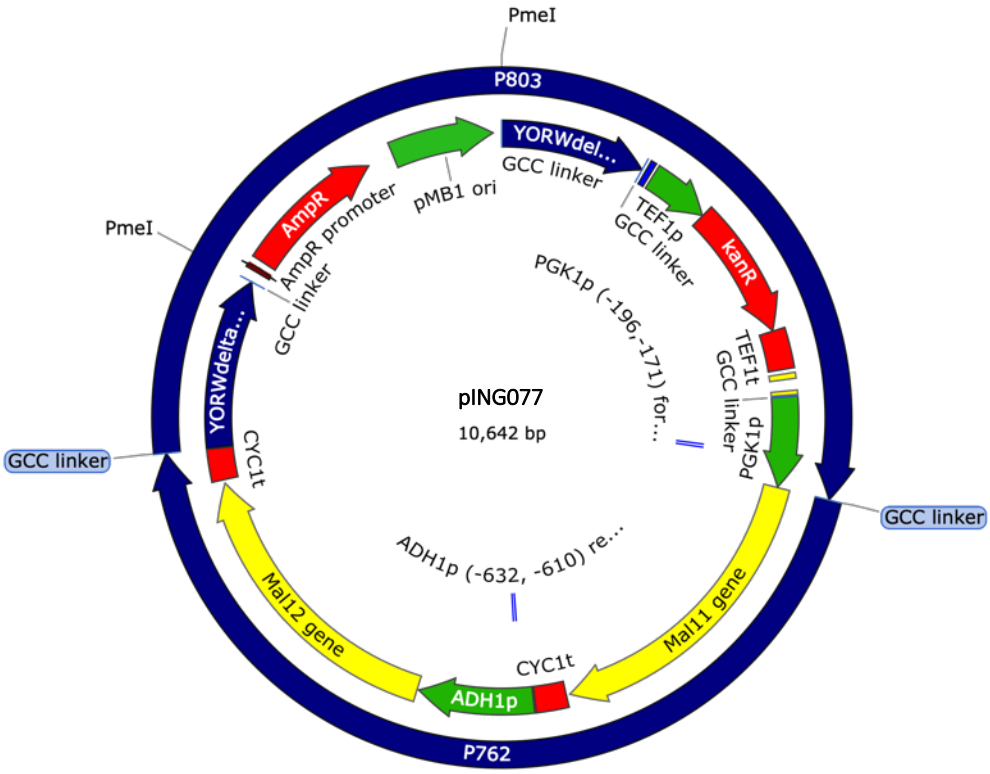


Figure 2.2. Plasmid map of pING077. Containing maltose transport and utilisation genes with ampicillin resistance gene for maintenance in *E. coli* and kanamycin resistance for selection in *S. cerevisiae*. pING077 contains the same sequence as pINGO76 apart from the inclusion of PKG1p instead of ADH1p for the MAL11 genes

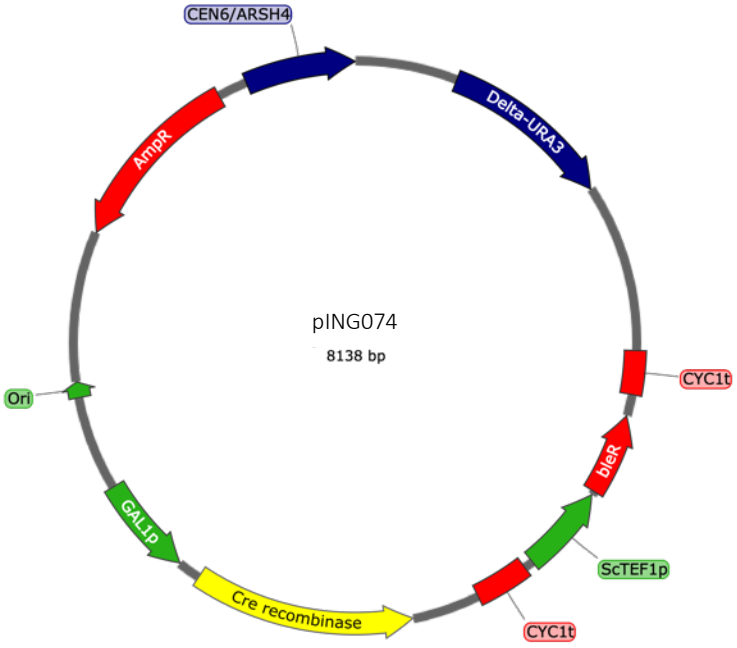


Figure 2.3. Plasmid map of pING074. Containing Cre recombinase and ampicillin resistance gene for maintenance in *E. coli*.

2.4 *S. cerevisiae* Fermentation

Fermentations were carried out using a DASGIP Parallel Bioreactor. Industrial strains Ethanol Red, Angel and laboratory strain S288c were streaked on a YPD plate from glycerol stocks and incubated for 48 hours at 30 ^{O}C. From YPD plates a single colony of each strain was inoculated to 5 ml YPD liquid medium and grown for 48 hours at 30 ^{O}C with 220 rpm shaking. From this culture 100 µL was inoculated to 50 ml liquid YPD culture and incubated under the same conditions as previously described.
Each vessel was prepared by filling with 4.5 liters of YPD media before being sealed and sterilised by autoclave. Once sterile, the chambers were attached to probes for pH and dissolved oxygen as well as ports for aeration, pH adjustments and a sampling port. All unused ports were sealed using a clamp and the system was allowed to equilibrate overnight.
The system was then set to 30 ^{O}C, with 100 rpm agitation. Ethanol Red and Angel were then inoculated into the DASGIP fermentation vessels, with each strain being inoculated into two five-liter vessels. The dissolved oxygen, CO_2 and optical density were measured over 27 hours.

2.5 Whole-genome Sequencing

2.5.1 DNA Extraction

2.5.1.1 *S. cerevisiae* short read

DNA was extracted using Yeastar Genome DNA kit from Zymo Research. (D2002) Yeast cultures were grown on YDP agar for 24 hours at 30^{O}C and from this culture single colonies were picked and inoculated into liquid YPD medium and grown overnight (~16 hours) at 30^{O}C with 200 rpm shaking (orbital shaker Thermo Scientific, Iowa, USA). DNA was extracted as per the manufacturer's guidelines as per protocol and the resulting DNA was eluted in 60 µL of nuclease free water.

2.5.1.2 *S. cerevisiae* optimised for long fragment length

To increase the fragment length of extracted DNA for Oxford Nanopore long read sequencing applications the recommended extraction kit was used. The Thermo Scientific Pierce Yeast DNA Extraction Kit(78870) was used in accordance with manufacturers guidelines. The resulting DNA was eluted in 50 µL of nuclease free water.

2.5.1.3 *Streptomyces* short-read DNA Extraction

Streptomyces strains were extracted using a phenol chloroform method as follows. Cells from a 5 ml overnight TSB culture of each strain were resuspended in 300 µL TE Buffer with 5 mg/ml lysozyme and 0.1 mg/ml RNase (Thermo Fisher EN0531) and incubated at 37^{O}C. 50 µL 10% Sodium Dodecyl Sulfate (SDS) (Thermo Scientific 28312)
Solution (was added and mixed thoroughly. 85 µL 5 M NaCL was added before mixing. 400 µL of phenol/chloroform/isoamyl alcohol (25:24:1) was added and the sample was vortexed for

30 seconds before centrifuging at 8000 rpm for 10 minutes. The aqueous phase was then transferred to a new micro-fuge tube and 500 µL isopropanol was added. This was mixed by inversion and left to incubate at room temperature for 5 minutes. The DNA was then pelleted at 10,000 rpm (Thermo Scientific, Megafuge 40R, USA) before being washed twice in 1 ml cold, 70% ethanol. All the ethanol was removed, and the sample was left open in a fume cupboard for >30 minutes or until completely dry. The pellet was then suspended in 50 µl – 200 µl of ddH_2O depending on required concentration.

2.5.2 DNA quantification

2.5.2.1 Nanodrop

The Thermo Scientific NanoDrop UV-Vis spectrophotometer 2000c (Thermo Scientific, USA) was used for quantification and quality checks by measuring 1 µL of DNA at 260/280 nm and 260/230 nm.

2.5.2.2 Qubit Fluorometer

In addition to the Nanodrop, the Qubit 2.0 Fluorometer (Invitrogen) was also used to measure DNA quantity, typically 10 µL of a 1/10 dilution using the high sensitivity dsDNA kit (Thermo Fisher Q32851). Each measurement was conducted in triplicate and an average was taken.

2.5.3 Library Preparation

2.5.3.1 Illumina Sequencing

Illumina Sequencing was carried out in-house using an Illumina MiniSeq system. Genomic DNA was prepared for sequencing with the Nextera DNA Flex library preparation kit (Illumina 20018704) and Nextera DNA CD indexes (Illumina 20018707) following the manufacturer's instructions.

2.5.3.2 Oxford Nanopore Sequencing

2.5.3.2.1 Library Preparation

Libraries of long fragment length DNA was prepared using the Ligation Sequencing Kit (SQK-LSK109) and Native Barcoding genomic DNA expansion pack (EXP-NBD104). DNA concentration was measured by Qubit and 1 µg of genomic DNA for each sample was made up to 49 µL with nuclease free water (W4502 Sigma-Aldrich). The DNA repair and end-prep step was carried out by combining 48 µL DNA, 3.5 µL NEBNext FFPE DNA Repair Buffer (M6630 New England Biolabs) , 2 µL NEBNext FFPE DNA Repair Mix (M6630 New England Biolabs), 3.5 µL Ultra II End-prep reaction buffer (E7546 New England Biolabs) and 3 µL Ultra II End-prep reaction mix (E7546 New England Biolabs) and incubated at 20°C for 5 minutes then 65°C for 5 mins using a thermocycler. 60 µL of thoroughly resuspended AMPure XP beads (A63881 Beckman Coulter) was added to the reaction mixture before it was shaken by hand for 10 minutes. The sample

was then placed on a magnet and the beads were washed twice with 200 µL of fresh 70% ethanol. After all ethanol was removed, the beads were allowed to dry for ~ 30 seconds. The beads were then resuspended in 25 µL nuclease free water and incubated at 37°C for 20 minutes. After incubation the beads are pelleted on the magnet and 25 µL of eluate was transferred to a 1.5 ml Microcentrifuge DNA LoBind tube (Z666548 Sigma-Aldrich) and quantified by Qubit.

2.5.3.2.2 Native Barcode Ligation

22.5 µl of end-prepped DNA (500 ng), 2.5 µL Native Barcode, 25 µL TA Ligase Master Mix were added in order stated with mixing after each addition. The reaction was incubated at room temperature for 10 minutes. 50 µL of resuspended AMPure XP beads to the reaction mix and incubated for 15 minutes at room temperature with gentle shaking. The samples were spun down and placed on the magnet until the supernatant was discarded. The beads are washed twice with 200 µL of fresh 70% EtOH. Any residual EtOH was pipetted off and the beads were allowed to dry. The tube was then removed from the rack and the pellet was resuspended in 26 µL of nuclease free water. The sample was then incubated at room temperature for 5 minutes with gentle shaking before placing back on the magnet until the eluate was clear. 26 µL was removed to a clean LoBind tube. 1 µL of the sample was taken for quantification by Qubit.
Equimolar amounts of each barcoded samples were pooled into a LoBind tube to give 700 ng of total DNA in 65 µL nuclease-free water.

2.5.3.2.3 Adapter Ligation and Clean up

65 µL (700 ng) of pooled sample, 5 µL of adapter mix II (AMII), 20 µL of NEBNext quick Ligation Reaction Buffer (5X) and 10 µL of Quick T4 DNA Ligase were combined and mixed before being incubated at room temperature for 10 minutes. 50 µL of AMPure XP beads were added and the sample was incubated at room temperature with gentle shaking for 10 minutes. The samples were placed on a magnetic rack and the supernatant was removed. The beads were washed in 250 µl of Long Fragment Buffer (LFB) twice. After washing the beads were returned to the magnetic rack and the supernatant removed. 15 µl of Elution Buffer (EB) was added to resuspend the bead pellet and the sample was incubated at 37°C for 45 minutes. The beads were then pelleted on the magnet and 15 µl of eluate was removed to a fresh LoBind tube. 1 µl of the sample was taken for quantification by Qubit.

2.5.3.2.4 Port Priming and Sample Loading

Air was removed from the priming port by dialling back on a P1000 until liquid entered the tip. 800 µl of Priming Mix was added to the flow cells via the priming port, carefully to avoid the introduction of air bubbles. The library was prepared by mixing 37.5 µl Sequencing buffer (SQB), 27.5 µl Loading Beads (LB) and 12 µl of DNA Library (up to 50 fmol). 200 µl of priming mix was added to the priming port and finally 75µl of sample was added to the SpotON port in a dropwise fashion. The SpotON port was closed and the MinION was attached via USB to a computer and data acquisition was started via the MinKNOW program.

2.6 Sequencing data analysis

2.6.1 Sequencing Analysis Workflow overview

The same pipeline was followed for all assemblies in the first instance with modifications made where necessary. These workflows are outlined in this section with more detail given to each tool subsequently.

2.6.1.1 **Short reads**

Illumina sequencing was carried out using the in-house Illumina MiniSeq. The primary data analysis method is outlined in Figure 2.4 with the boxes representing the analysis software and arrows representing the flow of data. Curved arrows represent the iterative nature of the process, sometimes required to ensure more accurate assemblies. Iterative rounds were performed until no improvement in the assembly was seen.

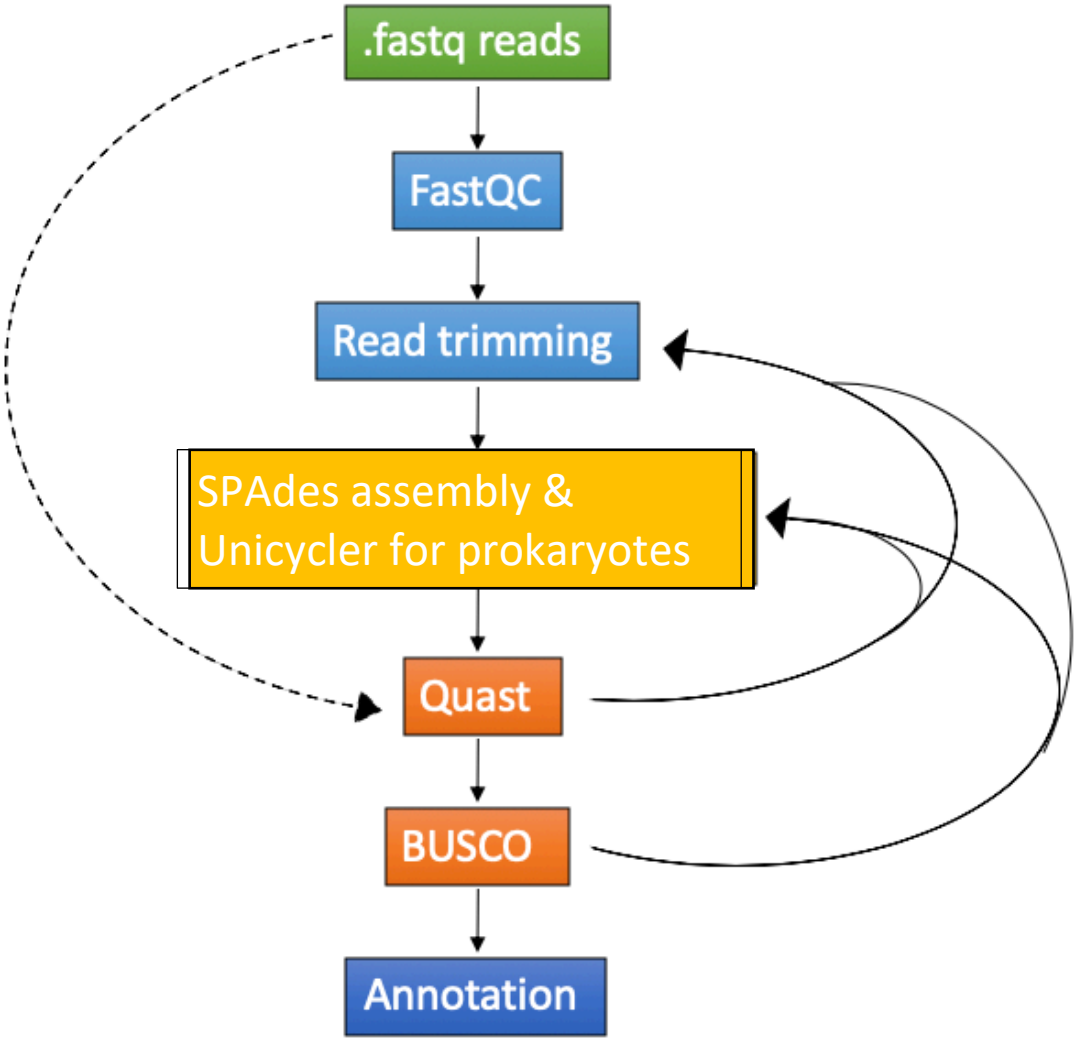


Figure 2.4. Sequencing analysis workflow. Showing the various analysis steps undertaken in the processing of short-read sequencing data from raw data input (green), though quality control (blue) assembly (yellow), assembly quality assessment (orange) and annotation (blue). Curved arrows depict where multiple iterations of certain parts of the process may be required to optimise parameters. The Dashed arrow indicates that reads can be mapped onto the final assembly to act as an additional quality control method.

2.6.1.2 **Long reads**

Long read assembly followed the same fundamental pipeline, with the exception being that the output from an Oxford Nanopore Technologies (ONT) MinION was in a different format to most short-read technologies and thus requires conversion to fastq. Format before further processing. The basic workflow for this is shown in figure 2.5.

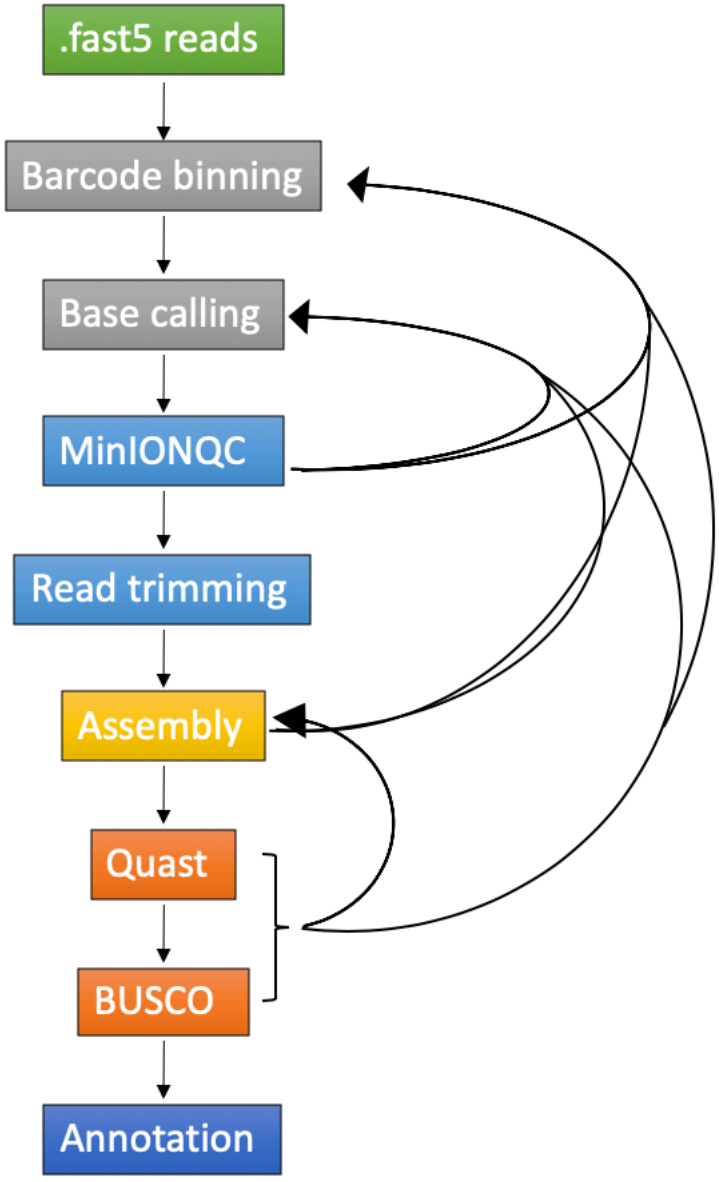


Figure 2.5. Sequencing analysis workflow. Showing the various analysis steps undertaken in the processing of long read sequencing data from raw data input (green), though quality control (light blue) assembly (yellow), assembly quality assessment (orange) and annotation (dark blue). Curved arrows depict where multiple iterations of certain parts of the process may be required to optimise parameters.

2.6.1.3 Hybrid **assembly**

Hybrid assembly uses both short and long read to provide the best assemblies. The general workflow for long read assembly that was used in this work was shown in Figure 2.6.

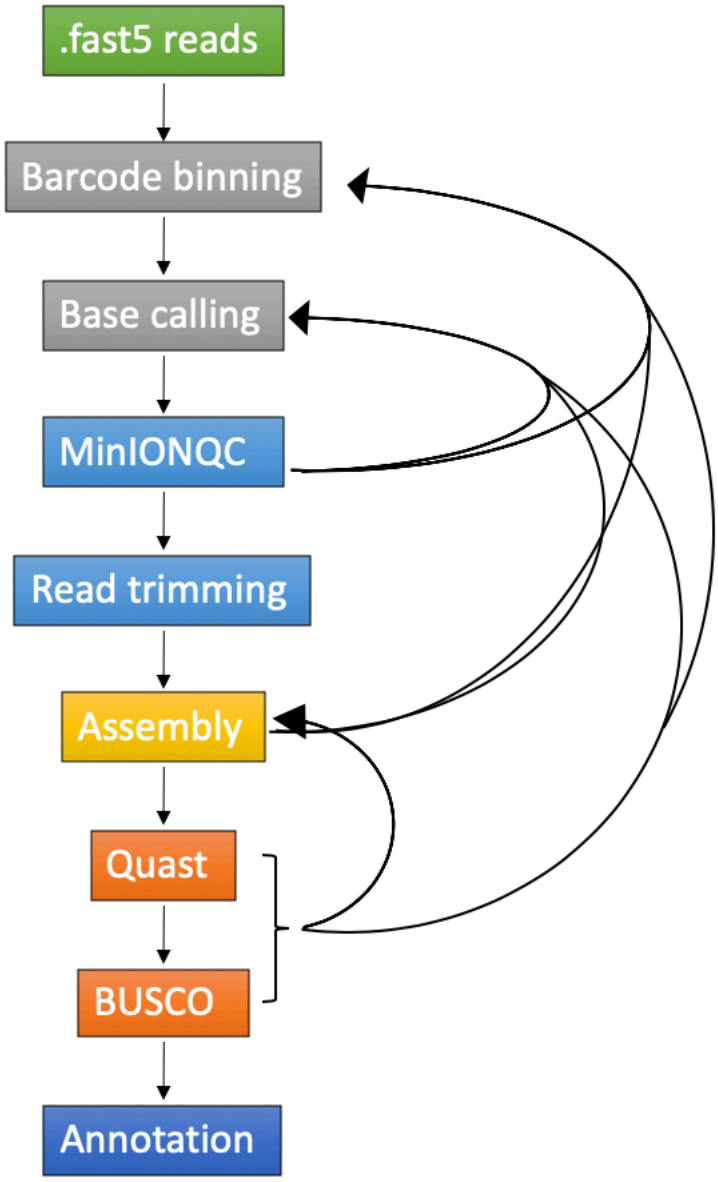


Figure 2.6. Sequencing analysis workflow. Showing the various analysis steps undertaken in the processing of long read sequencing data from raw data input (green), though quality control (light blue) assembly (yellow), assembly quality assessment (orange) and annotation (dark blue). Curved arrows depict where multiple iterations of certain parts of the process may be required to optimise parameters.

2.6.2 Raw Read Quality Assessment

2.6.2.1 Illumina Read QA

The read quality can be assessed by the Phred score which was attributed to each base during the sequencing process. Phred scores are calculated as the probability that a base was called incorrectly. This is determined by the functioning of the instrument at each particular base call, meaning that a high Phred score across all reads is indicative of accurate sequencing. Here the program FastQC (Andrews 2014) was used to assess the quality scores of all of the reads. FastQC graphs the read length quality as a function of read length to allow for further processing of the data as well as raw data statistics such as read length, GC content and adapted abundance. Based on this data some sets of reads were trimmed to maintain a high quality of data prior to assembly. Reads with an average quality of below 35 Phred were trimmed using Trimmomatic (Galaxy Version 0.36.5) in Galaxy EU, using sliding window trimming with an average over 4 bases.

2.6.2.2 MinION Read QA

To assess the quality of the MinION reads the R program MinIONQC (Lanfear *et al.*, 2019), which was available on GitHub at https://github.com/roblanf/minion_qc, was used. To perform this analysis the sequencing_summary.txt file produced by the guppy basecaller was used to produce various plots that allow for assessment of the reads on a variety of parameters including read length, mean Q score, read length as a function of Q score, length, quality and number of reads per hour, yield by length, yield over time and gb per channel. MinIONQC was run using the command: *Rscript MinIONQC.R -i path_to_file/sequencing_summary.txt -o /path/to/outputdir -f "png".*

2.6.2 Read Assembly

2.6.1 **De-novo Assembly**

2.6.1.1 SPAdes

Illumina reads were assembled using SPAdes v3.13.1 which uses a de Bruijn graph-based assembly algorithm, using the parameters: *spades.py -1 illumina_R1.fastq.gz -2 illumina_R2.fastq.gz –careful.* Where long reads were also used the *--nanopore /path/to/long_reads* option was also included.

2.6.1.2 Canu

For Nanopore-only assembly Canu was used. Canu was cloned from GitHub repository https://github.com/marbl/canu and was used with the following command:
-nanopore-raw /path_to_.fastq -p yeast03canu -d /path_to_file genomeSize=12m

2.6.2.3 Flye

The alternative method for Nanopore only assembly was Flye. It was cloned from GitHub repository https://github.com/fenderglass/Flye and used with the following command:
bin/flye --nano-raw /path_to_file/.fastq --genome-size 12m --out-dir /path_to_file/new --threads 40

2.6.2.4 Draft Based Scaffolding

MeDuSa (Bosi *et al.*, 2015) was used to produce a scaffolded assembly by aligning contigs to a reference genome. The input was the contig file output of which ever assembly tool was used with the .fasta of an assembled reference genome.

2.6.3 Assembly Assessment

QUAST was used to give an assessment of assembly quality (Gurevich *et al.*, 2013), the input for this was the FASTA assembly output from the assembly program used and the output was a PDF giving various metrics including total assembly length, N50, L50, GC%. QUAST was accessed from the internal server and run using the following command: quast.py --min-contig -o /path/to/output /path/to/.fasta.

BUSCO was also used to determine assembly completeness, this was used on the Galaxy.eu server, BUSCO completeness (Galaxy Version 3.0.2+galaxy2). Default parameters were used with --lineage_path set to *bacteria_odb9* for Streptomyces and fungi_odb9 *for S. cerevisiae.*

Ideel was cloned from the GitHub repository using: git clone https://github.com/phiweger/ideel and the Diamond index used was the Swiss-Prot reviewed database from UniProt.

2.6.4 Whole-Genome Alignment and Visualisation

Mauve was used to visualise the whole genomes of the strains assembled in this work as well as the reference genomes and to generate a comparison between genome/assembly structure. Progressive MAUVE was used which allows for regions of identity to be identified that are present in some but not all genomes in the alignment. From this alignment the "weight" was calculated based on the number of mismatches. The weight determined here is used to differentiate more and less sensitive (higher and lower weighted) alignments. The default was used here which was 3 times the minimum match size to define the Locally Collinear Blocks (LCBs). After defining LCBs MAUVE uses MUSCLE (Edgar, 2004) and ClustalW (Larkin *et al.*, 2007; Sievers and Higgins, 2014) to align each one.

Integrated Genome Browser (IGB) (Freese, Norris and Loraine, 2016) was used to view assemblies and produce visualisations of both short and long read coverage over the

combinatorial assemblies. Tablet (Milne *et al.*, 2016) and Artemis (Rutherford *et al.*, 2000) were also used for genome visualisation and browsing and manual curation of SNP and indel sites.

2.6.5 Combinatorial Assembly Improvement.

2.6.5.1 **Pilon**

In instances where both long and short-reads were obtained, the long reads were first assembled using Flye then Pilon was used to detect errors in this assembly using the paired end Illumina data. Pilon (Galaxy Version 1.20.1) was utilised on the https://usegalaxy.org.au server with a paired end fragment BAM file (produced using Bowtie2 also within Galaxy Version 1.20.1 to map forward and reverse Illumina reads to the assembly) provided against the .fasta file for the Flye assembly and the following settings: organism was diploid=yes default base quality=15 minimum depth=0.1. Pilon was run iteratively until the number of variants found of each strain against itself plateaued (Walker et al. 2014).

2.6.6 Phylogeny inference

In the case of the *Streptomyces* strains AutoMLST (Alanjary, Steinke and Ziemert, 2019) accessed at https://automlst.ziemertlab.com/analyze was used to determine the phylogenetically closest organism, to each of the four MED strains. Fast alignment mode (MAFFT FFT-NS-2) was used.

2.6.7 Secondary metabolite prediction

Since it was expected that the *Streptomyces* strains would have a considerable biosynthetic capacity AntiSMASH 5.0 (Blin, Shaw, *et al.*, 2019) was utilised to predict which secondary metabolite clusters were present in these genomes and determine their most likely products. AntiSMASH was accessed at https://antismash.secondarymetabolites.org/#!/start and all additional features were used (KnownClusterBlast = ON, ClusterBlast = ON, SubClusterBlast = ON, MIBiG cluster comparison = ON, ActiveSiteFinder = ON, RREFinder = ON, Cluster Pfam analysis = ON, Pfam-based GO term annotation = ON, TIGRFam analysis = ON) .

2.6.8 Whole Genome Annotation

Annotation of the assembled yeast genome was carried out first using the yeast genome annotation pipeline (YGAP) which annotates yeast strains based on identity to sequences found in the Yeast Gene Order Browser (YGOB)(Byrne and Wolfe, 2005; Proux-Wéra *et al.*, 2012). The parameters used for this were post whole-genome duplication species = yes, fix frameshift using estimates = yes and ordering scaffolds by size = yes.

Prokaryotic annotations were carried out using RAST (Rapid Annotation using Sub-system Technology) or PROKKA (Seemann 2014). RAST which is based on SEED (Aziz et al. 2008; Overbeek et al. 2014). For RAST annotations the parameters were as follows Genetic code: 11, Annotation Scheme: ClassicRAST, preserve gene calls: no, automatically fix errors: yes, fix frameshifts: yes, backfill gaps: yes.

2.6.10 **Whole Genome Sequence Based Comparison**

In each case assemblies were compared to each other initially by genome size and GC% using the metrics given by QUAST (Gurevich *et al.*, 2013) and the number of genes found by BUSCO. All assemblies were compared to each other and at least one reference genome on a whole-genome scale using Mauve (Darling *et al.*, 2004) and CCTMaps (Grant, Arantes and Stothard, 2012). Snippy was also used to determine the presence or absence of single nucleotide polymorphisms between strains.

2.6.10.1 CGView Comparison Tool

CGView Comparison Tool (CCT) maps were used to visualise and compare whole genomes. Based on the areas identified in these analyses as being of interest, further investigation of smaller regions was carried out using a variety of methods depending on the strain and type of analysis required. For comparison of aligned homologous genome regions the SEED viewer (Aziz *et al.*, 2008) and CLUSTAL (Larkin *et al.*, 2007; Sievers and Higgins, 2014) were used. NCBI BLAST (Johnson *et al.*, 2008; Sayers *et al.*, 2009) was used to find homologs in a custom BLAST databases that contained each of the sequenced genomes. The NCBI BLAST+ makeblastdb (Galaxy Version 0.3.3) function was utilised within the Galaxy.au platform with default parameters to generate both nucleotide and protein BLAST databases from the genome in which query sequences were being searched for, a FASTA file of the query sequence was generated to be searched against the database using megablast for nucleotide queries and Blastp for protein queries both with an expectation value cut-off of 0.001 (Camacho *et al.*, 2009; Cock *et al.*, 2015).

2.6.10.1 Snippy

The Snippy tool was used within (Galaxy Version 4.3.6+galaxy2) with the reference genome as the assembly for comparison and paired end reads of the second strain added and the settings were as follows: mapping quality – 60, minimum coverage – 10, - minumum proportion for variant evidence – 0.9, minimum quality - 100

2.7 Maintenance of *S. cerevisiae* Plasmids in *E. coli*

2.7.1 Plasmid Confirmation by Restriction Digest

Plasmids provided by Ingenza was digested using Pme1 (New England Biolabs) as follows.

Component	Volume

Pme1	1 µl
Cutsmart Buffer	10 µl
pING077 or pING076	1 µg
Nuclease free water	Up to 50 µl

The reagents were combined into a 1.5 ml Microcentrifuge tube and incubated on a heat block at 37°C with 250 rpm shaking for 1 hour. After incubation 5µl of the digested DNA was run on a 1% agarose gel for 1 hour at 80V.

2.7.2 Preparation of *E. coli* Competent Cells

A single colony of *E. coli* DH5α was inoculated to LB medium and incubated overnight at 37°C with 200 rpm shaking. 2.5 ml of this overnight culture was added to 250 ml of LB and incubated until the OD_{600} reading was between 0.5 and 0.6 (~3-4 hours). The culture was incubated on ice for 15 minutes and centrifuged at 1900 x g for 10 minutes at 4°C. The supernatant was discarded, and 2 ml cold sterile H_2O was used to resuspend the pellet. Cold sterile H_2O was added until the total volume was 200ml. The cell suspension was centrifuged at 1900 x g for 10 minutes at 4°C. The pellet was resuspended in 50 ml cold, sterile H_2O. The cell suspension was centrifuged once again with the same conditions and the pellet was resuspended in 1 ml of cold sterile water.

2.7.3 Transformation of Plasmids to *E. coli*

200 µl of competent cells are mixed with 3 µl of plasmid (~1µg) and electroporated in a Gene Pulser cuvette 0.1 cm electrode (BIO-RAD) in a BIO-RAD Gene Pulser set to 2.5 Kv, 400 Ohms, 25 µF. After electroporation the cells are added to 1 ml of LB and incubated for 1 hour at 37°C. After 1-hour recovery 100 µl of the culture it plated on LB + 100 µg/ml ampicillin.

2.7.4 Recovery of plasmid DNA from *E. coli*

Promega Wizard SV Miniprep kit was used as per manufacturer's instructions to purify plasmid from 5 ml of *E. coli* culture and the plasmid DNA was eluted in 50 µl of molecular biology grade water.

2.7.5 Restriction Digest of Plasmid

Purified plasmid from *E. coli* was digested using Pme1 (New England Biolabs) as follows.

Component	Volume
Pme1	8 µl
CutSmart Buffer	80 µl
pING077 or pING076	8 µg
Nuclease free water	Up to 400 µl

The reagents were combined in a 1.5 ml Microcentrifuge tube and incubated on a heat block at 37°C with 250 rpm shaking for 1 hour. After incubation the enzyme was inactivated by incubation on a heat block at 65°C for 20 minutes and then the digested DNA was run on a 1%

agarose gel for 1 hour at 80V. The desired band was excised and purified using the Promega Wizard SV PCR/Gel clean up kit following the manufacturers guidelines.

2.8 Plasmid Transformation to *S. cerevisiae*

5 ml of YPD was inoculated with a single colony and incubated overnight at 30°C with 200 rpm shaking. $OD_{600\ was}$ measured and the number of cells per millilitre was calculated, based upon the approximation that a suspension of $1x10^6$ cells/ml will give an OD_{600} of 0.1 and the relationship was linear. 2.5 x 10^8 were added to 50 ml of YPD medium in a 500 ml Erlenmeyer flask and incubated at 30°C with 200 rpm shaking until the cell density was at least $2x10^7$ cells/ml (about 3 hours). Cells are centrifuged for 1 minute at 3000 x g and washed twice in 1 ml of sterile water. 1 ml of cell suspension was then transferred to a sterile 1.5 ml Microcentrifuge tube and centrifuged at 13000 x g for 2 minutes. The cell pellet was then resuspended in 1ml of sterile water and aliquots of 10^8 cells were transferred to fresh Microcentrifuge tubes and were centrifuged at 13000 x g for 2 minutes. For each transformation the following reagents were combined to make the transformation mixture; 240 µl PEG 3500 (50% [w/v]), 36 µl LiAc (1M), 50 µl single stranded salmon sperm DNA (2 mg/ml, prepared by boiling for 5 minutes at 100°C),34 µl pING074 (1 µg). The transformation mixture was used to resuspend cells by vortexing for 30 seconds. The cells plus transformation mix were incubated in a water bath at 42°C for 38 minutes. After incubation the cells were pelleted, the transformation mix was removed, and the cells were washed in 500 µl of sterile water. The washed cells are resuspended in 1 ml YPD media and transferred to a 15 ml falcon tube to be incubated horizontally at 30°C with 200 rpm shaking for 2 hours. 100 µl of transformed cells was plated on YDP + 500 µg/ml G418 and incubated at 30°C for 48 hours.

2.8.1 Confirmation of Insertion

2.8.1.1 PCR Based Confirmation

Primers “KanR_fwd” and “KanR_rev” were designed to anneal to kanamycin resistance gene in the plasmid sequence. PCR was carried out as per manufacturer’s guidelines with MyTaq DNA polymerase 1% agarose gel for 60 minutes at 100 V.

2.8.1.2 Sequence Based Confirmation

The PCR product from the PCR based confirmation was purified using a Promega Wizard PCR/Gel Clean up Kit following the manufacturers guidelines. The DNA concentration was then quantified by Nanodrop and diluted with molecular biology grade water to 5 ng/µl in 15µl. This was prepared for both forward and reverse reaction for each sample and 2 µl of the appropriate primer at a concentration of 10 µM was added to the tube. These were then sent to Eurofins Genomics LTD for Sanger sequencing via the Tube Seq service.

2.8.2 Curing of *S. cerevisiae* Integrations

Single colonies were picked into 0.5 ml YPD-G418 and incubated overnight at 30°C with 200 rpm shaking. The OD_{600} of the culture was measured and 5.0 x 10^7 cells were added to 10 ml

of pre-warmed YPD liquid in an Erlenmeyer flask and incubated at 30°C with 200 rpm shaking until the cell density was greater than or equal to 2.0 x 10^7 cells/ml (around 3 hours). The cells were harvested by centrifugation at 3000 x g or 5 minutes and the cells were washed twice in 25 ml of sterile water before being resuspended in 1 ml sterile water. The cell suspension was then centrifuged at 15,000 x g and the supernatant was discarded. The cells were resuspended in 1 ml sterile water and aliquots of 10^8 cells were transferred to fresh 1.5 ml microcentrifuge tubes. For each transformation the following reagents were combined to make the transformation mixture; 240 µl PEG 3500 (50% [w/v]), 36 µl LiAc (1M), 50 µl single stranded salmon sperm DNA (2 mg/ml, prepared by boiling for 5 minutes at 100°C),34 µl pING074 (1 µg). The transformation mixture was used to resuspend cells by vortexing for 30 seconds. The cells plus transformation mix was incubated in a water bath at 42^OC for 38 minutes. After incubation the cells were pelleted, the transformation mix was removed, and the cells were washed in 500 µl of sterile water. The washed cells were resuspended in 1 ml YPD media and transferred to a 15 ml falcon tube to be incubated horizontally at 30°C with 200 rpm shaking for 3 hours. 100 µl of transformed cells was plated on YDP + Zeocin and incubated at 30°C for 3-4 days.

2.9 Omnilog: Biolog Phenotypic Microarrays

2.9.1 Plate Preparation

Biolog Plates were obtained from Omnilog premade *via* Technopath. The inoculating fluids were prepared as described in the manufacturer's guidelines. The *S. cerevisiae* being tested were inoculated to YPD medium overnight. After ~24 hours of growth at 30°C a single colony was taken and sub-cultured to a new YPD plates and grown again for 24 hours at 30°C. Cells from this plate were taken and suspended in sterile water to a transmittance of 62%. The plates were prepared as follows using a 96-channel pipette and covered with BreathEasy Membranes to reduce evaporation.

PM Stock Solution	PM 1, 2 (ml)	PM 3-8 (ml)	PM9+ (ml)
IFY-0 (1.2x)	20	60.0	-
SC Medium (1.2x)	-	-	70.0
Dye mix D (75x)	0.32	0.96	-
Dye Mix E (100x)	-	-	0.84
D-Glucose (24x)	-	3	3.5
Sterile water	3.18	6.54	9.19
Cell Suspension (48x)	0.5	1.5	1.75
Total	24.0	72.0	84.0

2.9.2 Biolog Data analysis

There are multiple analysis tools in existence such as OPM for R and Biolog's own software, but none could achieve a comparison of all growth conditions in way that is both detailed and easy to read. heatmaps.2 from gplots was utilised to plot the difference in growth at each timepoint

for two strains in each well in a more compact manner than traditional growth curves can achieve. The code used to perform this analysis can be found in a the R.biolog.R script at https://github.com/Ainsleydmb/biolog_analysis. To visualise the data heatmaps.2 was used to plot a comparison between two strains growing on the same condition by calculating the difference between normalised values from two plates. This allows differences in growth between strains in any one condition to be viewed in a heatmap format. In addition, the heatmaps.2 software groups the growth profiles based on their similarity which means that conditions were grouped based on whether both strains grow similarly, or if one outperforms the other. This code can also be found in GitHub as **R_biolog_compar.R** at https://github.com/Ainsleydmb/biolog_analysis/blob/master/R_biolog_compar.R

Chapter 3. Analysis of Industrially Relevant Yeast Strains

3.1 Introduction

3.1.1 Relevance of Yeast Genome Sequencing to industry

Whole-genome sequencing has become invaluable to industrial biotechnology where it can increase the volume and specificity of data on strains and facilitate targeted and streamlined experimental protocols. In addition, genomic information can be utilised by industry to protect and identify their intellectual property.

3.1.1 Current Reference Strain Availability

The two industrial strains investigated to be used as starting points from which to derive an industrial lineage were Ethanol Red and Angel. Of these Ethanol Red has a draft genome sequence publicly available, and Angel does not. The draft genome of Ethanol Red was published in 2015 by Lund University after sequencing using Illumina MiSeq with Accession number: SAMN03266447 and ID: 3266447 (Wallace-Salinas *et al.*, 2015) the summary statistics of this assembly are were given in table 3.1. Angel, which has no published genome available was sequenced by Edinburgh Genomics on behalf of Ingenza, but this data proved to be unreliable upon analysis, due to low quality and a lack of metadata leading to uncertainty that all reads were generated from the same strain. Three sets of forward and reverse reads were obtained, seemingly deriving from different sources referred to here as Angel, No_I and CGAT. FastQC quality graphs showing the lower-than-expected Phred scores for Illumina data were given in figure 3.1. These graphs show that the quality of the data declines towards the end of the 100 bp reads but that it is also of a low quality throughout, particularly in the reverse reads. It can also be seen from these graphs, given the difference between the median and the minimum that there could be an improvement made by trimming reads based on quality, however, even after quality trimming using Trimmomatic the quality of the assembly resulting from this data was questionable. There were three sets of sequencing reads provided by this sequencing effort and so they were each assembled separately as three different assemblies and also by combining all reads together to generate a single assembly to assess the quality of each data set. Table 3.2 summarises the resulting assemblies from SPAdes, each assembly is smaller in size than the S288c reference genome but with a large contig number ranging from 200 to > 2500 contigs. As such it was decided that in-house sequencing should also be carried out for Angel as well as Ethanol Red in order to obtain suitable quality reference genomes.

Metric	Result Ethanol Red
Total sequence length	11,492,342 bp
Coverage	301.0 x
Number of scaffolds	218
Scaffold N50	95,106 bp
Scaffold L50	40 bp
Number of contigs	223
Contig N50	94,962 bp
Contig L50	41

Table 3.1 Summary statistics of the currently available Ethanol Red genome sequence

Statistic	Angel	No_I	CGAT	Combined
# contigs (≥0bp)	60251	966	855	49177
# contigs (≥1000bp)	625	279	229	2636
Total length (>=0bp)	8927033	11564013	11702645	12821468
Total Length (≥1000bp)	1064794	11410610	11534296	10188859
Reference	12071326	12071326	12071326	12071326
GC %	40.64	38.27	38.1	38.07
Reference GC%	38.30	38.30	38.30	38.30
N50	1647	40851	132555	5210
# N's per 100kbp	0	0	0	0

Table 3.2 Summary statistics as determined by QUAST. Analysis of the assemblies of the various sets of reads provided for Angel (read set one = "Angel", read set two = "No_I" and read set 3 = "CGAT" with the default contigs included being those above 500 bp in length

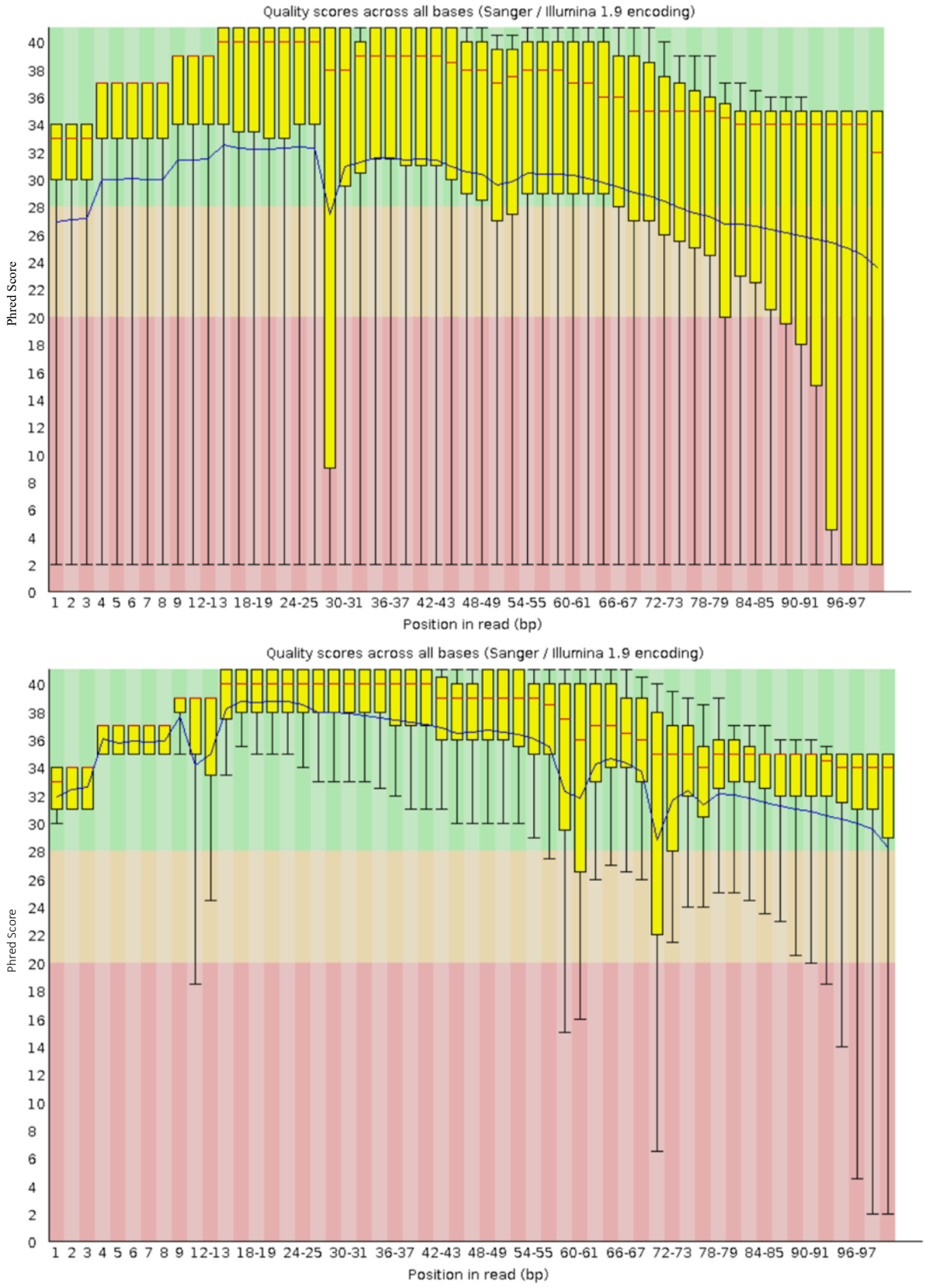


Figure 3.1. Raw Data Quality Check. FastQC results from the Edinburgh Genomics sequencing effort of Angel. A) gives the Angel reverse reads and B) shows the Angel forwards reads. Each box plot gives the median, minimum, maximum and 1st and 3rd quartile Phred score (Y-axis) for the data set at that base position (X-axis).

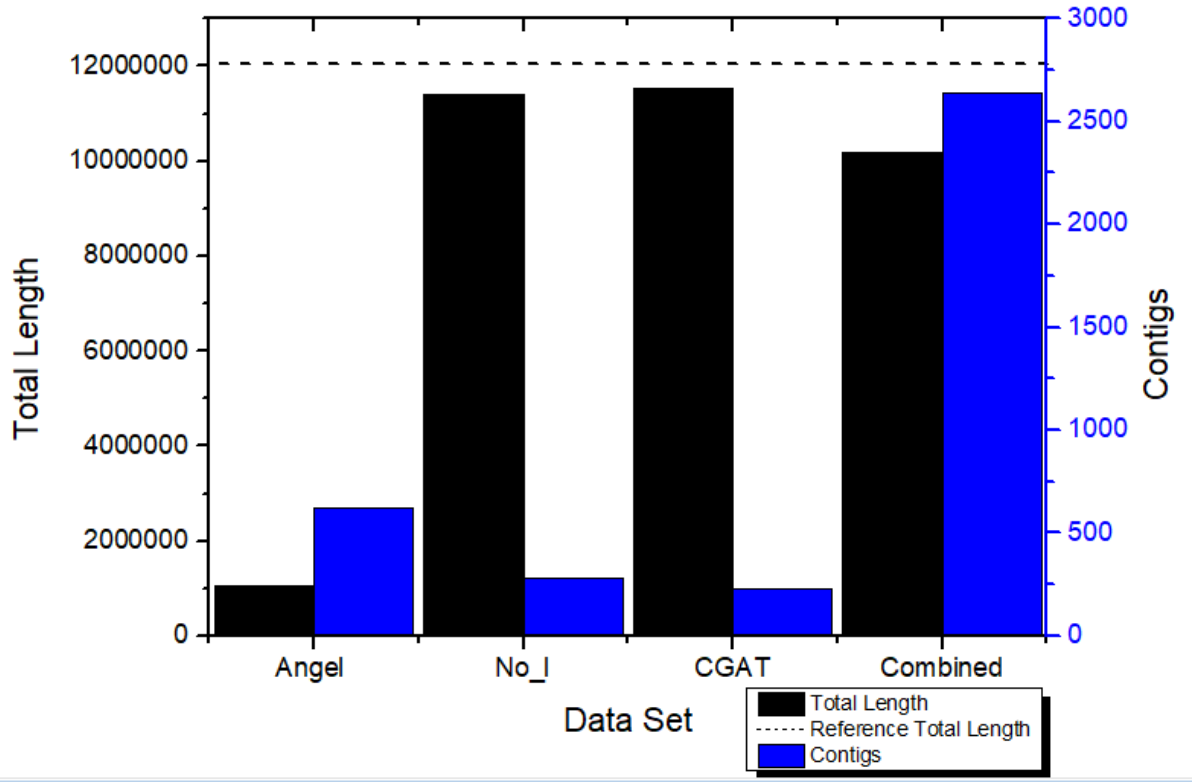


Figure 3.2. Assembly statistics. Shows the total length of each assembled genome (black bars) with the size of the reference genome indicated by a dashed line. The number of contigs in each assembly are shown in blue bars.

3.1.3 Aims and Objectives of this Chapter

In order to increase understanding of the adaptations that have already been made to Ingenza's *S. cerevisiae* strains for bioethanol production and what may be possible in the future, whole genome sequences were obtained for several industrially-adapted strains. This data was to complement pre-existing published data for the reference strain S288c, for which we also have phenotypic data. As previously discussed, the available sequences for reference strains Ethanol Red and Angel provided suboptimal references for further assemblies and a less than perfect comparison assembly for genomic analysis. As such the first main objective for this part of the work was to obtain better-quality reference genomes, and subsequently use this to improve the assemblies of all strains derived from these. Once this was achieved the second main objective was to use these data to analyse and determine the genomic differences between strains as well as to infer genomic influence over phenotype. In-house Illumina MiniSeq sequencing was used to provide high quality, short-read data that could be augmented with Oxford Nanopore MinION data as required to generate a bank of high quality-draft or complete yeast genome.

A deeper understanding of the genetic background to the strains currently under investigation by industrial partners Ingenza will allow for future modifications to be informed. This will also allow for differences between the strains, that may not be evident in laboratory conditions, to be factored into future strain selection. Furthermore, the information provided by a comprehensive whole genome sequence is vital for industrial quality control and patenting purposes.

3.2 Results and Discussion

3.2.1 Illumina Sequencing

3.2.1.1 Reference Strains

The Illumina MiniSeq was used in-house with the Nextera DNA flex library prep kit with native barcoding to sequence both Angel and Ethanol Red genomes.

3.2.1.1.1 Quality Assessment

FastQC (Andrews, 2014) was used to assess the quality of the raw reads and the results of this are given in figure 3.3. These data, as assessed by Phred score, are of higher quality than the previously obtained sequencing data for Angel in that the higher average Phred score over the length of the reads is higher, indicating a higher confidence that each base was correctly determined during sequencing.

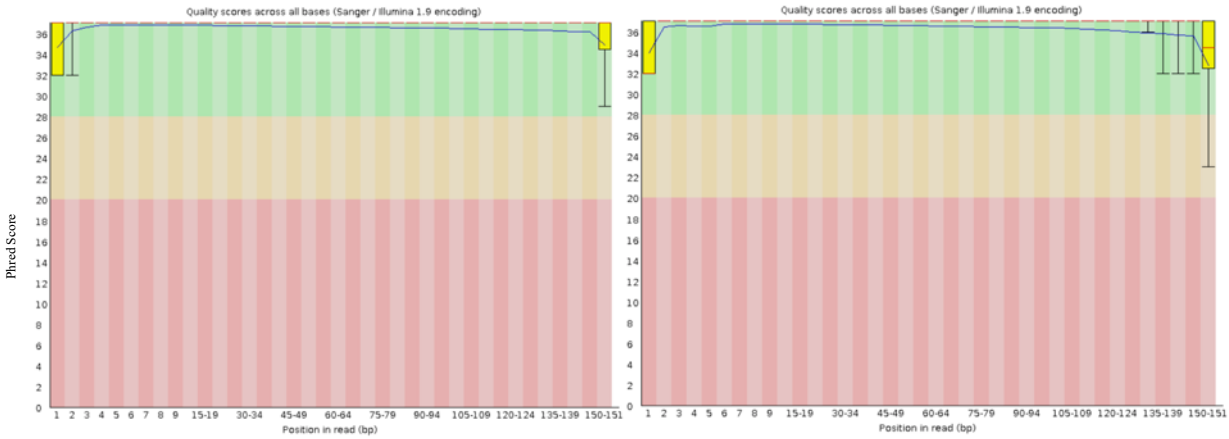


Figure 3.3 Raw Data Quality Check. Showing the FastQC results of the in-house Illumina sequencing being of a high quality and not requiring trimming before assembly. The left graph shows the forward reads and the right shows the reverse.

3.2.1.1.2 SPAdes Assembly

SPAdes (Bankevich *et al.*, 2012) was used to assemble both Ethanol Red and Angel genomes with default settings to provide draft level assembly status. The summary statistics obtained by QUAST analysis are given in table 3.3.

Statistics	Angel	Ethanol Red
# contigs	891	265
# contigs (>= 0 bp)	3140	900
# contigs (>= 1000 bp)	685	223
Largest contig	331762	362755
Total length	11511721	11542328
Total length (>= 0 bp)	11911286	11651030
Total length (>= 1000 bp)	11366081	11513050
N50	65164	123041
N75	16498	66356
L50	48	33
L75	133	65
GC (%)	38.09	38.09

Table 3.3 Summary statistics as determined by QUAST. Analysis of assemblies of the two industrial reference strains Angel and Ethanol Red with the default contigs for included being those above 500 bp in length unless otherwise stated.

3.2.1.2 Industrial Lineage

Six strains, known as 93, 94, 132 133, 134 and 135, that represent an industrial lineage, shown in figure 3.4, initially derived from Ethanol Red were also sequenced by Illumina MiniSeq to allow for single nucleotide polymorphisms as well as larger genetic differences.

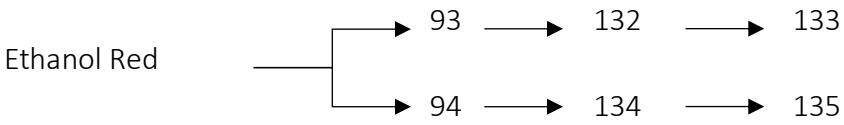


Figure 3.4. Industrial lineage progression. Schematic depicting the relationship between the industrial candidate strains of *S. cerevisiae* according to the order in which they were developed at Ingenza.

3.2.1.2.1 Quality Assessment

FastQC (Andrews, 2014) was used to assess the quality of the raw reads and the results of this are given in figure 3.5.

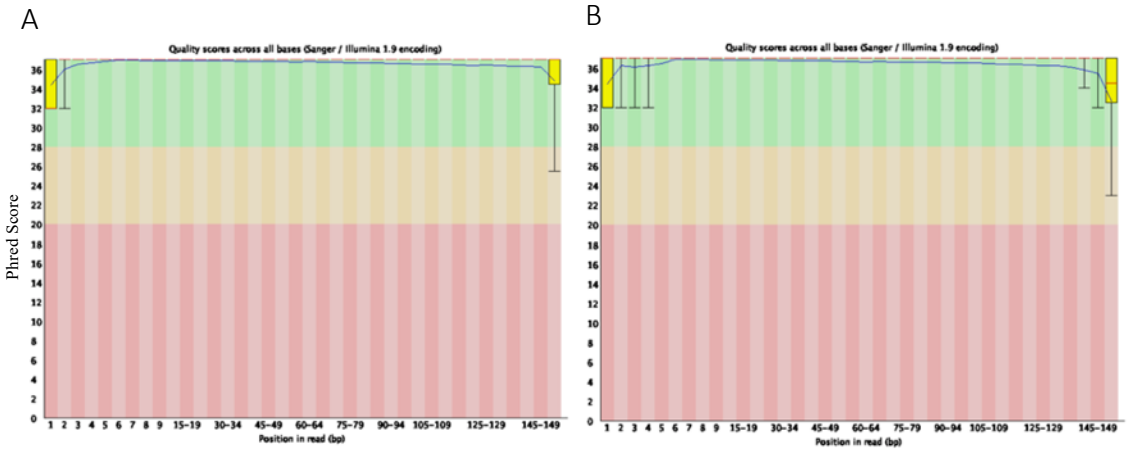


Figure 3.5 Raw Data Quality Check. Showing selected FastQC results of the in-house Illumina sequencing being of a high quality and not requiring trimming before assembly. A) Forward reads from 93. B) Reverse reads from 93.

3.2.1.2.2 SPAdes Assembly

SPAdes (Bankevich *et al.*, 2012) was used to assemble all of the genomes in the lineage with default settings to provide draft level assemblies. The summary statistics obtained by QUAST analysis are given in table 3.4. This shows that these assemblies are in a large number of contigs which is suboptimal for further analysis as there will be a large potential for information to be lost from between contigs as well as the presence of genes truncated due to an artefact of these assemblies.

Statistics without reference	93	94	132	133	134	135
# contigs	4543	3611	5107	3489	3430	3819
# contigs (>= 1000 bp)	980	796	1028	781	782	827
Largest contig	257930	258616	295805	314300	314300	348971
Total length	11496814	11518352	11489325	11530609	11511608	11497515
Total length (>= 1000 bp)	11259224	11340024	11194213	11360904	11367772	11337592
N50	51050	54319	50747	55066	51796	51674
L50	59	54	57	52	53	58
GC (%)	38.09	38.09	38.09	38.09	38.09	38.1

Table 3.4 Summary statistics as determined by QUAST. Analysis of the assemblies of the industrial lineage with the default contigs for included being those above 500 bp in length unless otherwise stated.

3.2.2 Nanopore Sequencing

3.2.2.1 Quality Assessment

To assess the quality of the Oxford Nanopore MinION reads, MinIONQC was used within RStudio to determine the average read quality and output metrics for the entire data set. The pores within a MinION sequencer are grouped into channels, and MinIONQC measures the average result per channel in each case. Figure 3.6 gives the read abundance over a range of quality scores and indicates that a large proportion of the total reads were of a high quality. Figure 3.7 gives a better understanding of the run by showing the per-channel quality score, this indicates that the quality decreases with time as the pore becomes more used and as such the longer running of the MinION may not be beneficial. This also indicates that almost all channels provide high quality reads at the beginning of the run and continue to do so for a large proportion of the run time. In addition, the channels that produce more low-quality reads, or no reads at all are not clustered together and this indicates proper loading of the MinION with no bubbles. Figure 3.8 gives various graphs relating to the quality and output of the MinION run over the 48-hour time span. Figure 3.8 A gives the yield as a product of time and shows that the majority of the yield is obtained between hours 0 and 32 with output tailing off after this point. This graph displays this metric for both total yield and the yield of reads with quality score of above 7 only and this shows that the proportion of low quality reads also increases towards the end of the run, however, the drop in yield when only high-quality reads are considered is relatively low for this sequencing method. Figure 3.8 B. gives the mean quality score over time which corroborates the data in the last graph discussed as the quality does decrease over time from being at a mean of > 10 for quality trimmed reads until hour 40 before a slight drop. The 'all reads' section shows a slightly more pronounced dip in mean quality but remains over 7.5 even at its lowest. Figure 3.6 C. shows the mean length over the time course of the run. This starts low as any longer reads that are in pores at the start of the run take longer to be processed and as such don't appear as sequenced reads until later which explains the short increase in read length at the very beginning of the run. After around 15 hours the read length begins to tail off which can be explained by a combination of there being a limited number of long DNA fragments available for sequencing, DNA potentially being degraded in the MinION and the increased likelihood for pores to become damaged. This decrease however, is not sharp and so does not indicate any contamination by nucleases or significant issues with the MinION pores. Finally, figure 3.8 D. gives the number of reads per hour and this shows a steady decline over the entire course of the run. This trend is to be expected due to the inherent property of the MinION chemistry where pores become exhausted over time and the concentration of DNA available to be bound by sequencing pores reducing. In the instance where a the library was loaded on the MinION a second time, which can increase overall output, one would expect to see a spike on this graph shortly after the second loading. Finally figure 3.9 shows the output per channel across the MinION device and the distribution shown here indicates good device loading and even spread of DNA fragments throughout the pores. Poor device loading would have resulted in areas of no or low yield where air bubbles had been introduced and/or decline in yield towards the end of the device where loading speed or DNA input quantity was incorrect.

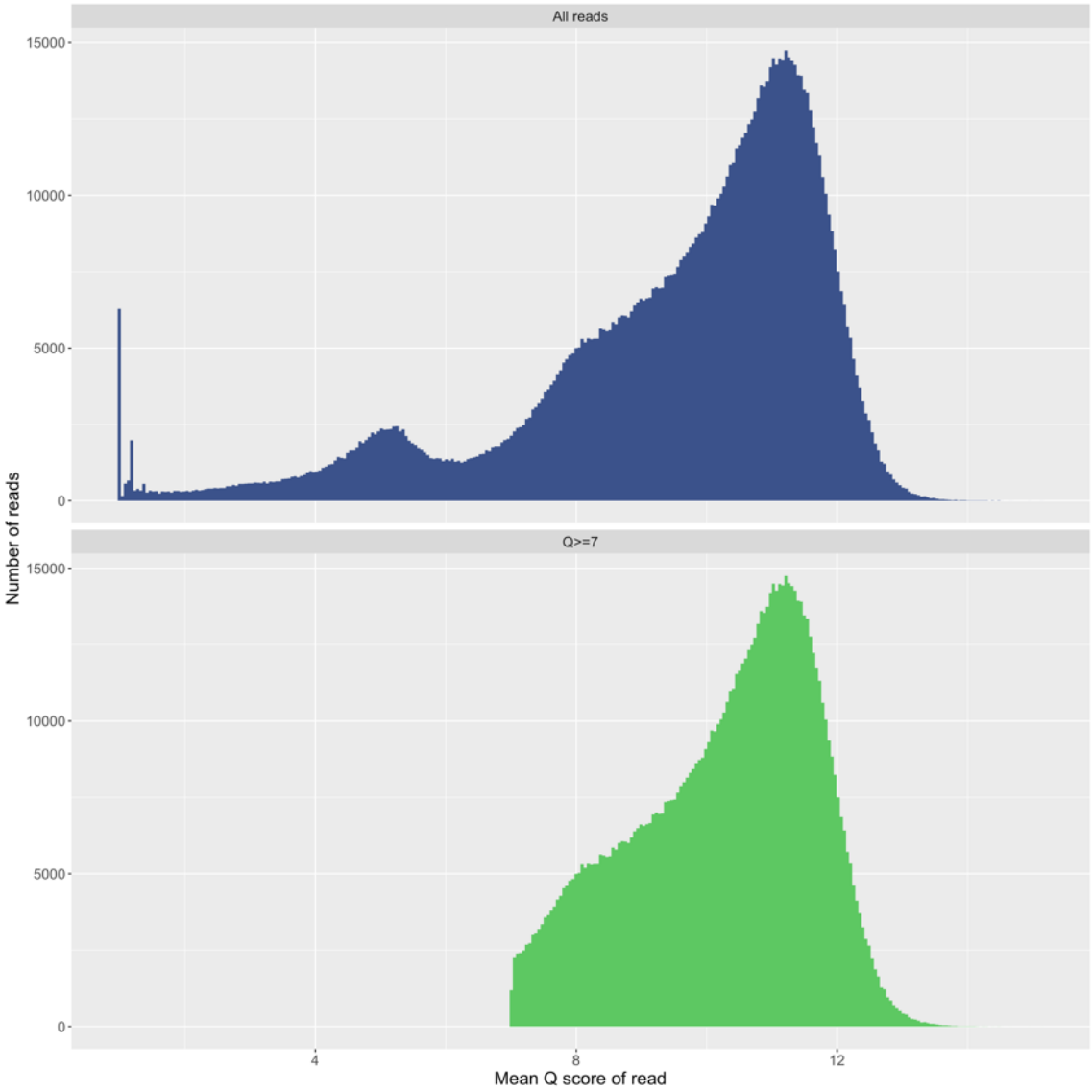


Figure 3.6. MinION Quality Control. Showing the mean quality score of MinION reads. All reads shown in blue with only reads with a quality of >7 shown in green

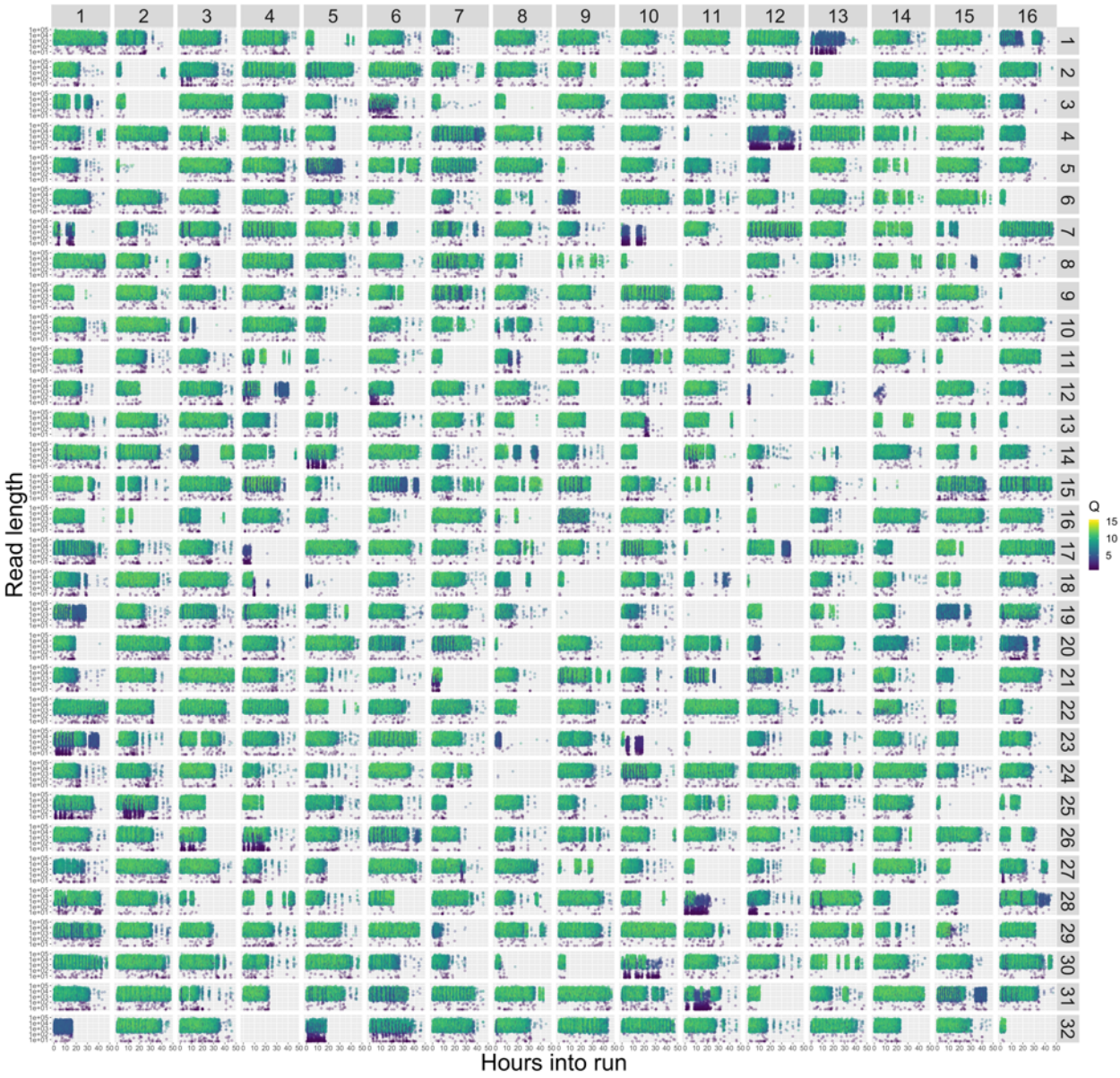


Figure 3.7 MinION Quality Control. Showing the quality score of MinION reads per channel as the run time increases from left to right for each channel. Quality decreases as the colour progresses from light green to dark blue.

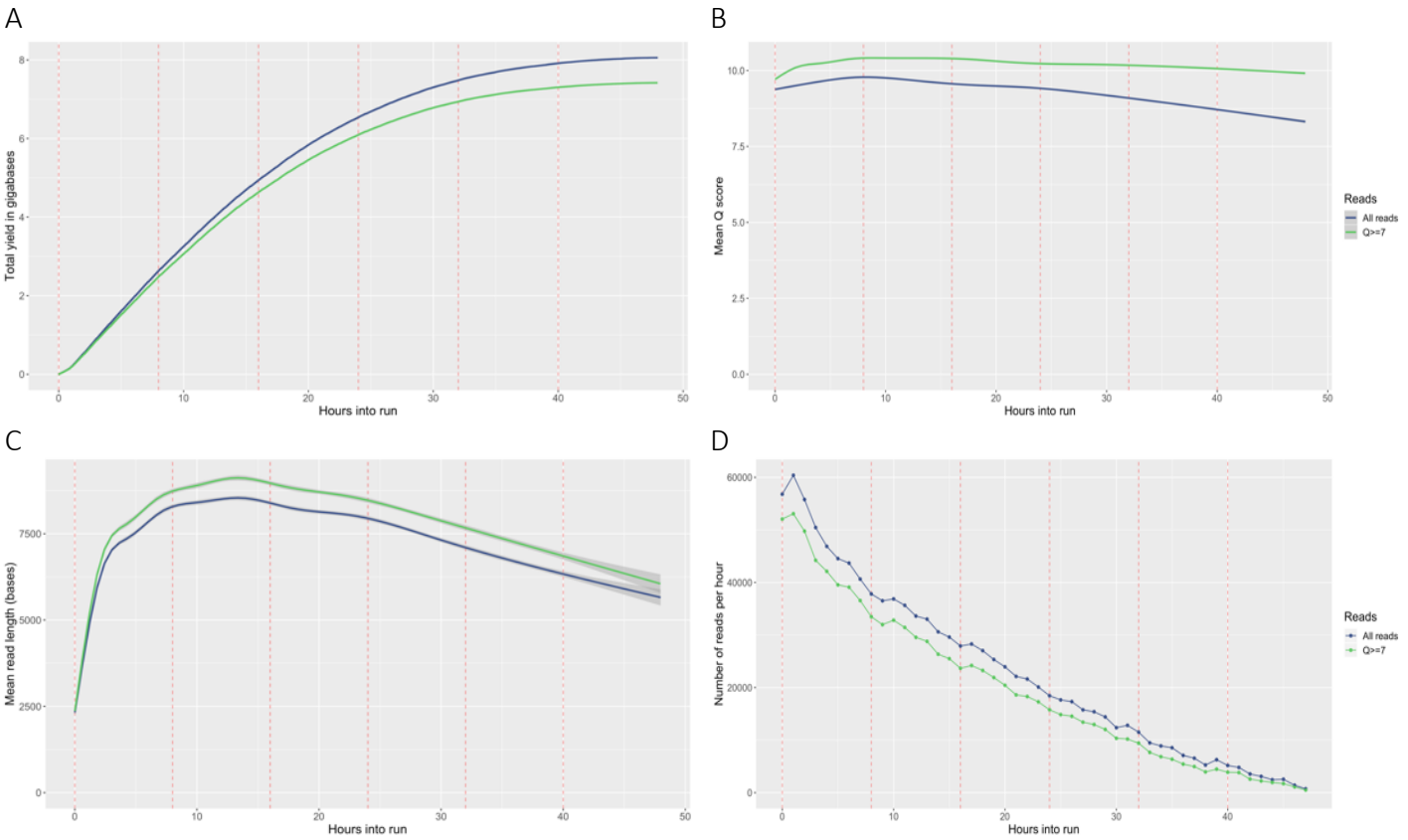


Figure 3.8. MinION Quality Control. A. Showing the sequencing yield in gigabases of MinION reads. In all cases all reads are shown in blue with only reads with a quality of >7 shown in green. B. Showing mean quality score of all reads as the run time progresses. C. Showing the mean read length with increasing run time. D. Showing the output of the sequencing run in terms of reads per hours as run time progresses.

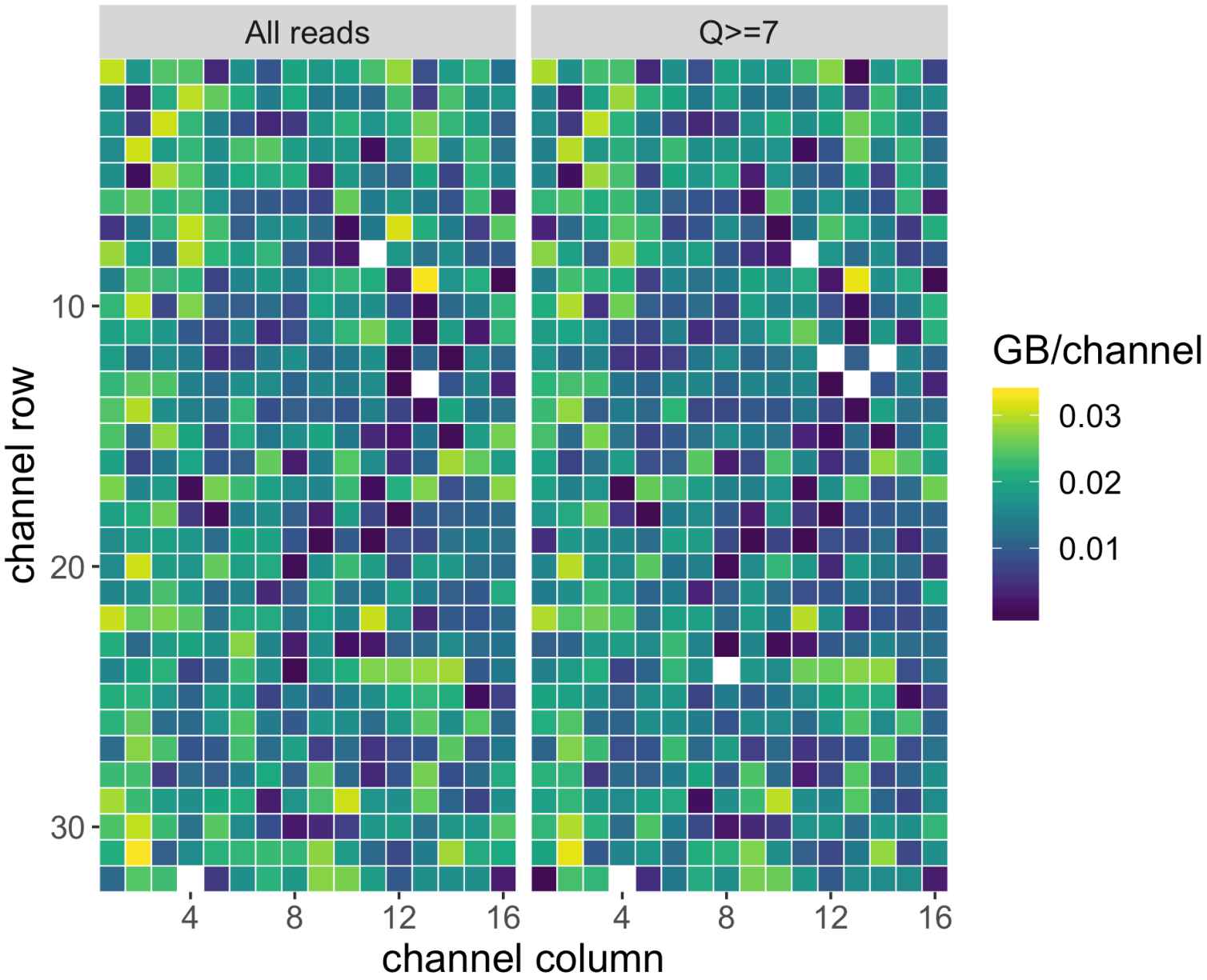


Figure 3.9 MinION Quality Control. Showing the output of each channel in gigabases. All reads shown in the left panel with only reads with a quality of >7 shown in the right panel.

3.2.2.2 Long Read Assembly

Nanopore reads were assembled with Flye (Kolmogorov *et al.*, 2019) and corrected by mapping short reads for the same strain obtained by in-house Illumina sequencing onto the Nanopore assembly using Pilon in order to provide a high quality assembly. The results of this assembly in comparison to the currently published Ethanol Red genome are given in table 3.5 with the contig sizes compared to that of the completed reference genome for the lab strain S288c being represented in figure 3.10.

Ethanol Red Assembly Statistics	Published assembly	Flye corrected with Pilon
# contigs	218	17
# contigs (>= 1000 bp)	218	17
Largest contig	442570	1472829
Total length	11492342	12066027
Total length (>= 1000 bp)	11492342	12066027
N50	95106	880186
L50	40	6
# N's per 100 kbp	2	0.08
GC (%)	38.1	38.09

Table 3.5 Assembly statistics table. Shown are statistics as determined by QUAST of the assemblies of the industrial lineage with the default contigs for included being those above 500 bp in length unless otherwise stated.

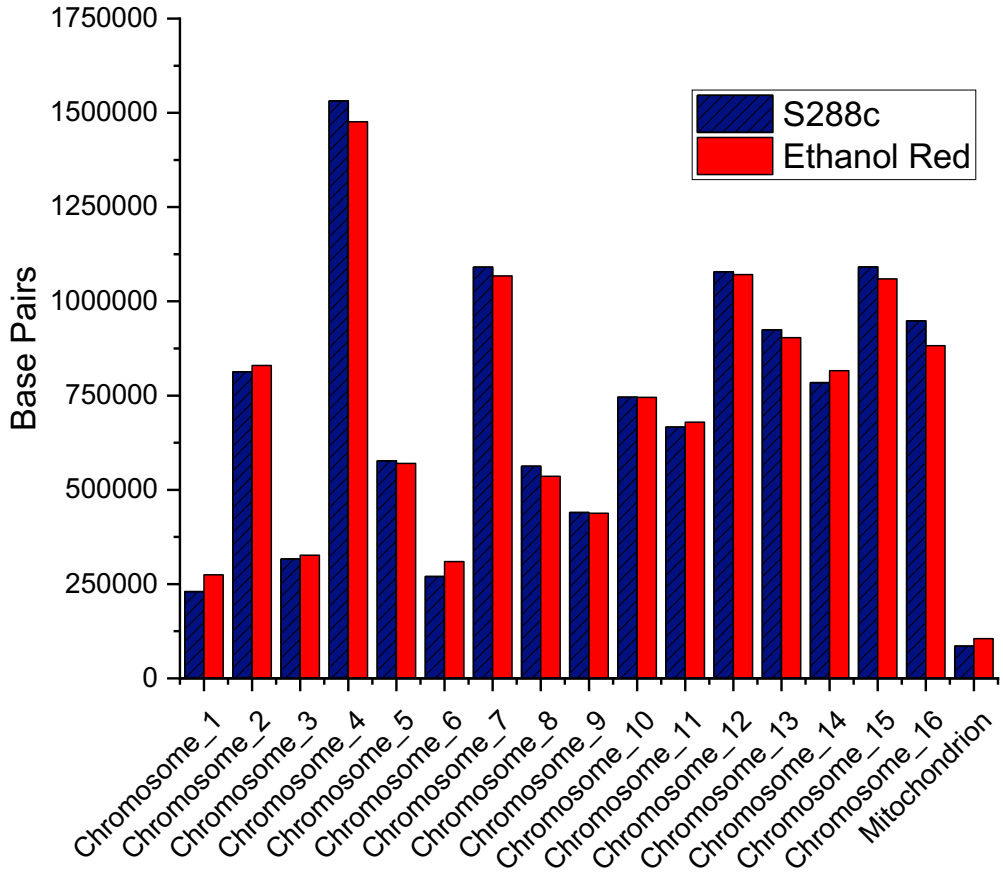


Figure 3.10. Ethanol Red Assembly Summary. Showing the length of each scaffold of Ethanol Red after scaffolding by MeDuSa to the S288c reference.

Long Read Assembly Quality

In addition to the QUAST assessment shown above, to investigate the Ethanol Red Nanopore assembly BUSCO and Ideel were run on the assembly. Ideel shows the number of proteins that are predicted to be truncated when compared to the manually-reviewed UniProt database, Swiss-Prot. Results from Ideel are shown in figure 3.11. These show that the short read correction of the Flye assembly has considerably reduced the proportion of highly fragmented genes in the Ethanol Red genome. Figure 3.11 A is the Flye assembly without correction, here there are 15033 genes <0.9 the length expected. Figure 3.11 B and C show the same assembly post Pilon assembly with 1233 fragmented proteins, indicating a significant improvement to the final assembly. Despite this improvement the final genome does still contain a significant number of truncated protein coding sequences. This means that there must be additional investigation before any conclusions could be drawn based on the presence/absence of protein coding sequence.

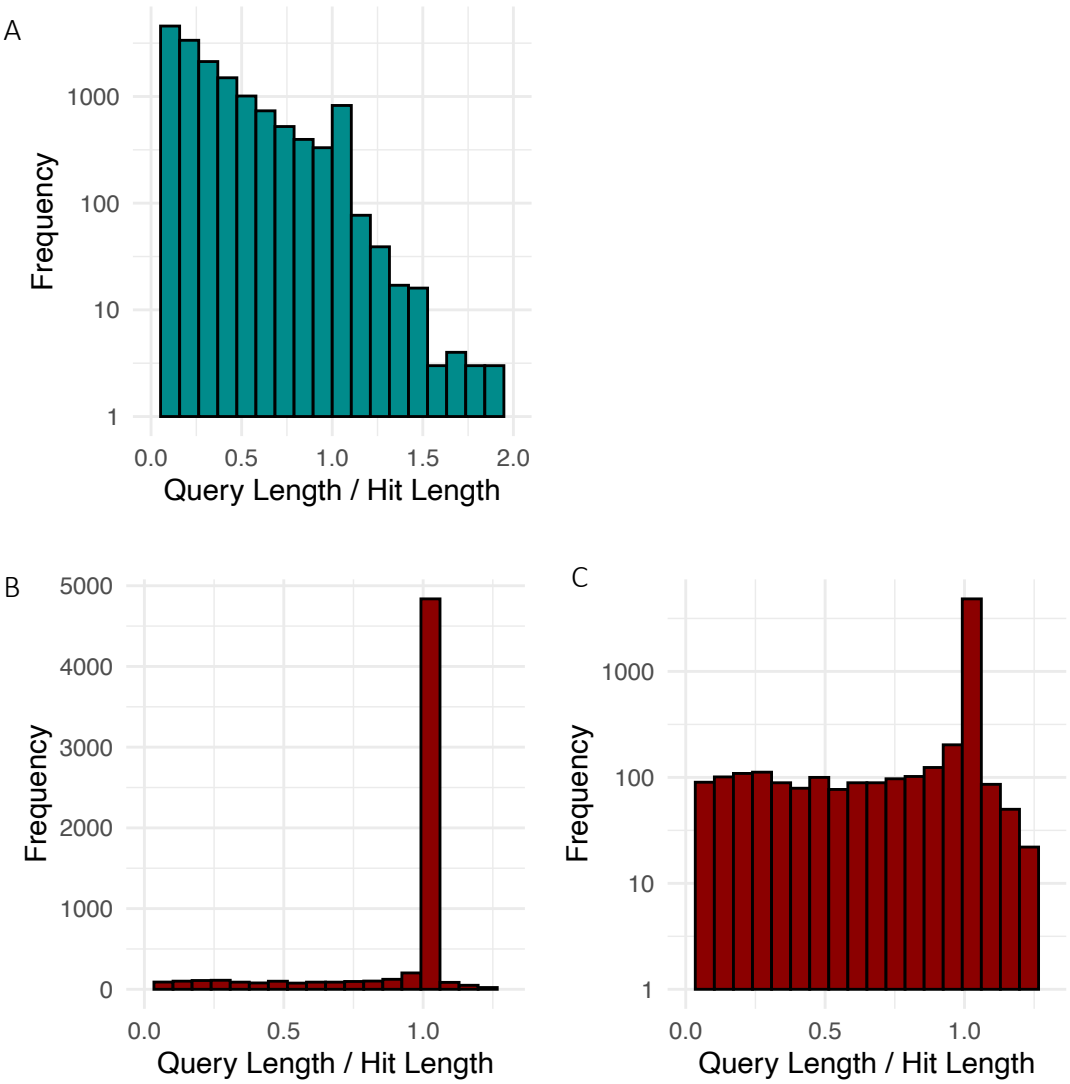


Figure 3.11. Histograms depicting Ideel protein length. Showing 15033 and 1233 proteins that are less than 0.9 of the total predicted length in the uncorrected and corrected assemblies respectively. Figure 3.9 A) is the uncorrected Flye assembly with 15033 genes <0.9 the length expected B) gives the total output of the corrected assembly with most of the proteins at 100% total predicted length but with a long tail of proteins <90% of predicted size. Figure 2.9 B) is the same graph scaled to show the lower end of the y-axis.

3.2.2.2.1 Reference-Guided Assembly

Due to now having a high-quality reference genome that can be used to improve assemblies, the complete Ethanol Red genome (GenBank: GCA_001078105.1) was used as a scaffold for these assemblies to improve their quality. This provides a scaffold for the assembly and helps determines a likely order and orientation for contigs of other strains. This resulted in a scaffolded genome as can be seen in table 3.6. Four out of six of these assemblies are in 17 scaffolds which is suggestive of 16 chromosomes and the mitochondrial sequence, this is corroborated by the GC% split between the chromosomes that represents the lower GC% content of the mitochondria. An example of this is shown in the GC% per contig graph generated by QUAST for strain 94 in figure 3.12.

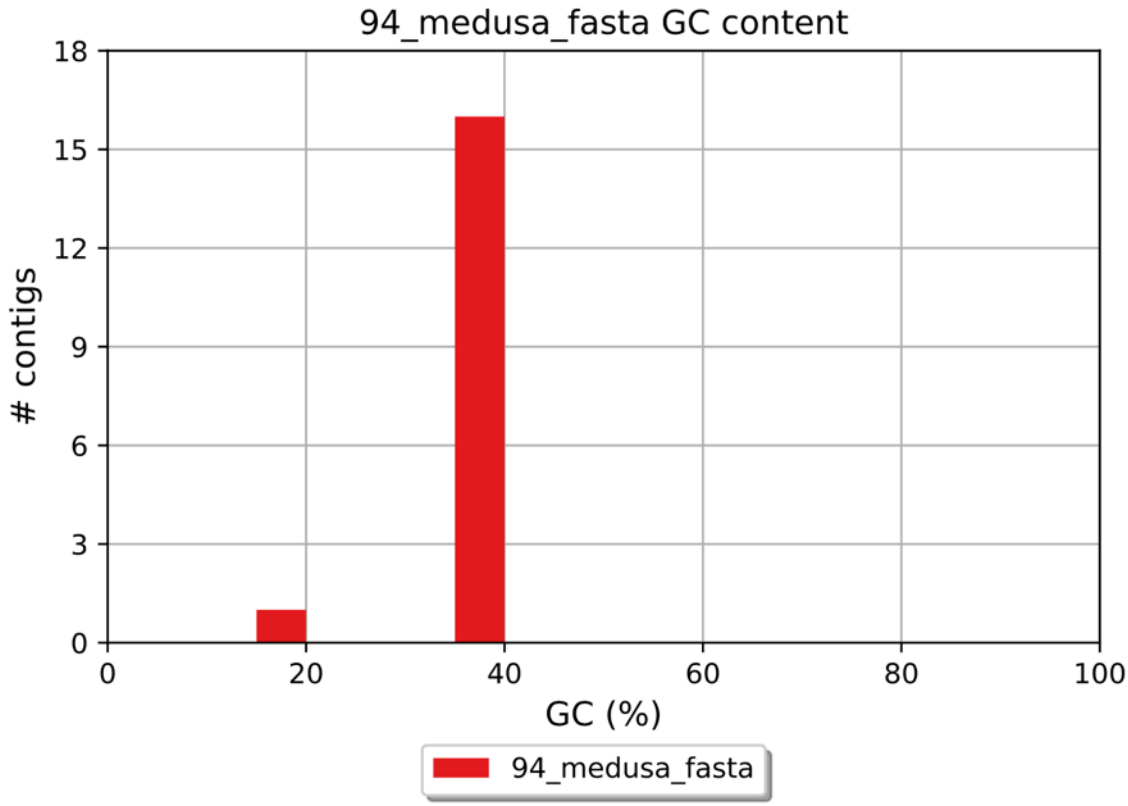


Figure 3.12. *S. cerevisiae* assembly GC%. Showing the QUAST GC% per contig graph with 1 contig at just <20% GC and the remaining 16 contigs just < 40% GC.

Statistics post MeDuSa	93	94	132	133	134	135
# contigs	848	734	995	675	782	782
# scaffolds (>= 1000 bp)	17	17	17	17	18	18
Largest scaffold	1593688	1427565	1434206	1596701	1588259	1588259
Total length	12136810	12004309	12201778	11997490	12021276	12021276
Total length (>= 1000 bp)	12136810	12004309	12201778	11997490	12021276	12021276
N50	879948	878958	880818	876922	881698	881698
L50	6	6	6	6	6	6
# N's per 100 kbp	1529.23	1192.07	1699.76	1149.41	1232.81	1232.81
GC (%)	38.1	38.1	38.1	38.1	38.1	38.1

Table 3.6 QUAST analysis of combinatorial assemblies. The statistics as determined by QUAST of the assemblies of the industrial lineage with the default scaffold included being those above 500 bp in length unless otherwise stated.

3.2.2.2.2 Reference-guided Assembly Quality Assessment

The qualities of the assemblies generated by the hybrid assembly approach were assessed for assembly completeness using BUSCO (Simão *et al.*, 2015). The results of this are given in table 3.7 and they show that there is a high degree of completeness over all strains. The missing BUSCO's are likely to present in the telomeric regions and this will be assessed and discussed in more detail. To further investigate the assembly quality Ideel was used to determine the proportion of truncated proteins that are predicted in each assembly, these figures are shown with the BUSCO results in table 3.7 and in graph format in figure 3.13. These suggest that whilst these results do show that most of the proteins are of the correct size there are still a significant proportion of truncated of or otherwise potentially misassembled proteins.

Strain Identifier	93	94	132	133	134	135
Complete BUSCOs (C)	2086	2095	2092	2093	2097	2096
Complete and single-copy BUSCOs (S)	2052	2058	2056	2059	2063	2062
Complete and duplicated BUSCOs (D)	34	37	36	34	34	34
Fragmented BUSCOs (F)	7	7	8	6	5	5
Missing BUSCOs (M)	44	35	37	38	35	36
Total BUSCO groups searched	2137	2137	2137	2137	2137	2137
Ideel, genes at <0.9 of predicted lenth	574	982	1164	1705	979	2417

Table 3.7. BUSCO analysis of industrial lineage. For each strain in the lineage the number of single-copy orthologues that are detected in each combinatorial assembly is given.

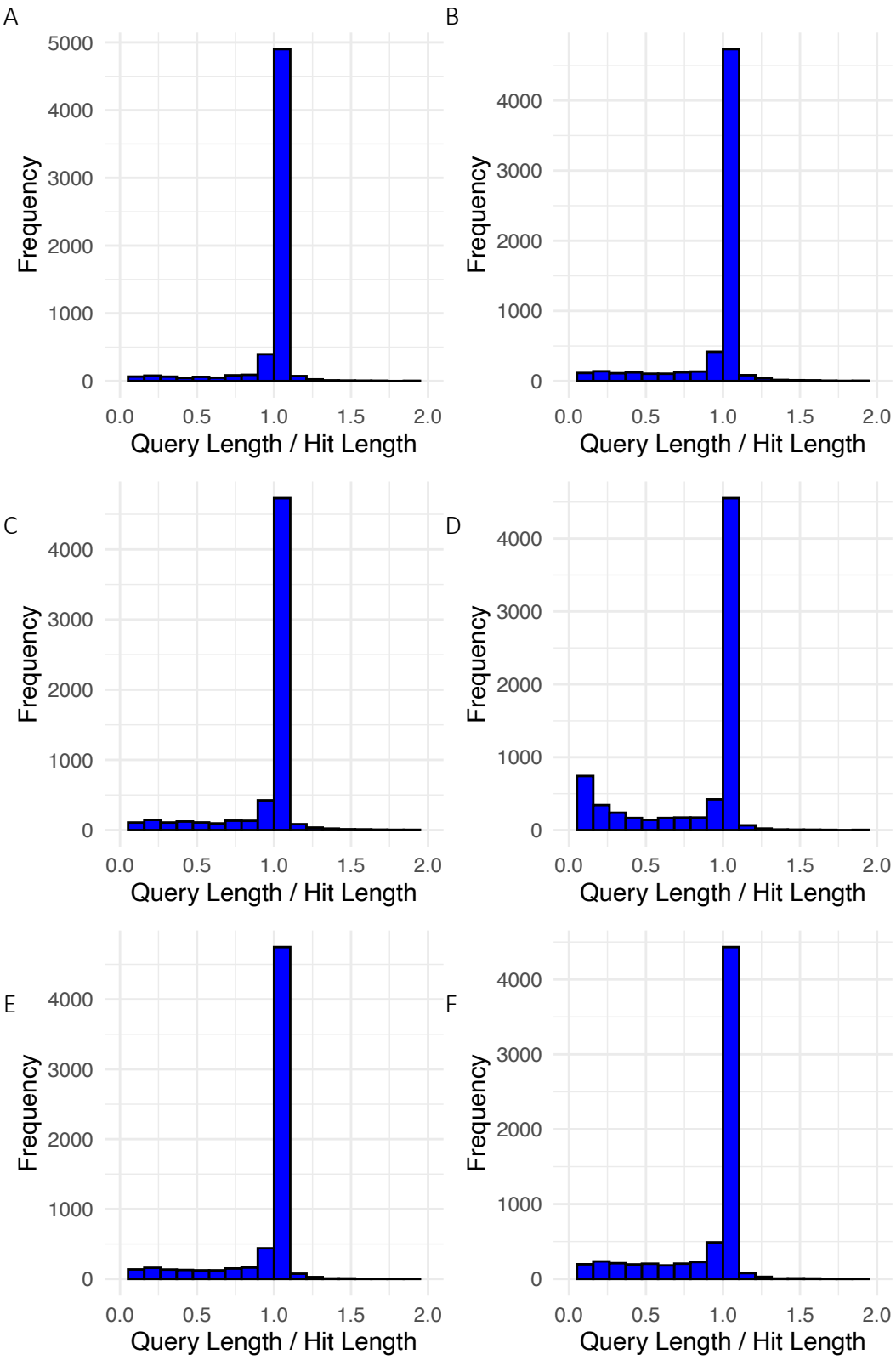


Figure 3.13 Ideel histogram. Output from Ideel in the form of histograms showing the relative sizes of predicted proteins as a ratio of the expected size in 93, 94, 132, 133, 134 and 135 are shown in A-F respectively.

3.2.3 Whole Genome Annotation

After final assemblies were generated for each of the strains RAST and YGAP annotation pipelines were utilised to annotate the genomes. Both were used as RAST provides a useful comparison tool for assessment of predicted ORFs whilst YGAP has been trained on an *S. cerevisiae* data set and so is most applicable to these genomes and as such provides a high degree of accurate annotations.

3.2.4 Sequence-Based Comparison of Yeast Genomes along an Industrial Lineage

After a high-quality reference genome had been obtained in the format of Ethanol Red, it was possible to use that for comparison of the 6 *S. cerevisiae* genomes that were generated by Ingenza as candidate strains for industrial bio-ethanol production. The lineage was generated by sequential rounds of adaptation beginning from Ethanol Red.

3.2.4.1 Comparison of Whole-genome Architecture

Initial comparison between the genomes was completed using CGView Maps. CCT maps were created for each chromosome and the mitochondrial sequence individually using the Ethanol Red genome as a reference with the six sequenced yeast strains as comparison genomes, ordered as per the sequence similarity of that chromosome to the corresponding reference chromosome. This similarity is determined by CGVeiw by using the scores of the blast hits as well as the coverage of these hits. The order in which the sequences are ordered for each chromosome varies, meaning that for each strain, its similarity to the Ethanol Red genome differs between chromosomes. This would indicate a potential error or experimental artifact if these strains were the result of natural evolution and variation as it would be expected that variation would occur more uniformly throughout the genome. Though, due to the nature of these strains, having been engineered in various ways including the insertion of heterologous genes, this pattern is not immediately indicative of an issue as significant selection pressure has been applied to these strains. Whilst each map shows a different pattern of similarity some traits are common amongst most or all chromosomes and the mitochondria. However, there is a difference in the similarity-based order between the CDS vs CDS maps and the DNA vs DNA maps. Many of the gaps in identity compared to Ethanol Red are common among all strains, which suggests that the six strains in the lineage are more like each other than they are to the reference, though as is previously observed there are truncated protein coding regions in the assembly and many of these differences occur at telomeric regions which are typically hard to assemble. Meaning that it is also possible that these apparent similarities are due to areas that are more prone to miss-assemblies. There is also a perceived gap in high percentage similarity (black regions) observed at the beginning and end (top of CCT map) of each sequence which likely corresponds to the reduction in coding sequences found at the telomeres in addition to varying levels of telomere sequencing completeness.

The map for chromosome one given in figure 3.14 depicts one of *S. cerevisiae*'s smaller chromosomes, with fewer predicted protein-coding regions. In this case, there appears to be a large degree of regions of this chromosome where no similarity is seen between the reference and comparison genomes, these areas largely correlate with regions with no predicted coding regions, in the CDS vs CDS maps gaps in predicted coding regions on the reference genome appear as a gap in similarity on the comparison genomes. Taking this into account there are only two areas containing coding sequences where there is low or no similarity between all six strains in the lineage and the reference. This is emphasised in the DNA vs DNA map for the same chromosome given in figure 3.15 where only two significant regions of difference can be observed that correspond to the regions of the CDS vs CDS maps that contain predicted coding regions on the reference genome. As such there are the two areas that are of the most interest. Zoomed maps were created of these two regions (figure 3.16 a and b). The zoomed maps show regions where 5 predicted protein-coding regions in Ethanol Red have low or no similarity in any of the six sequenced genomes. These regions are TETR0P00690 (SSA1), TETR0P01010, TETR0P01020, TETR0P01030, TETR0P01040, TETR0P01050 (Hypothetical proteins). SSA1 is a heat shock protein of the HSP7 family, with involvement in nuclear localization signalling (Shulga *et al.*, 1999). The read mapping of strain 94 to Ethanol Red shows a sharp drop in mapped reads at this location, which is suggestive that there has been a misassembly. To confirm this a blast search found the SSA1 in the mitochondria which would suggest a misassembly of these reads to the mitochondria. Also, possible but less likely is that there has been a translocation of these gene in the strain.

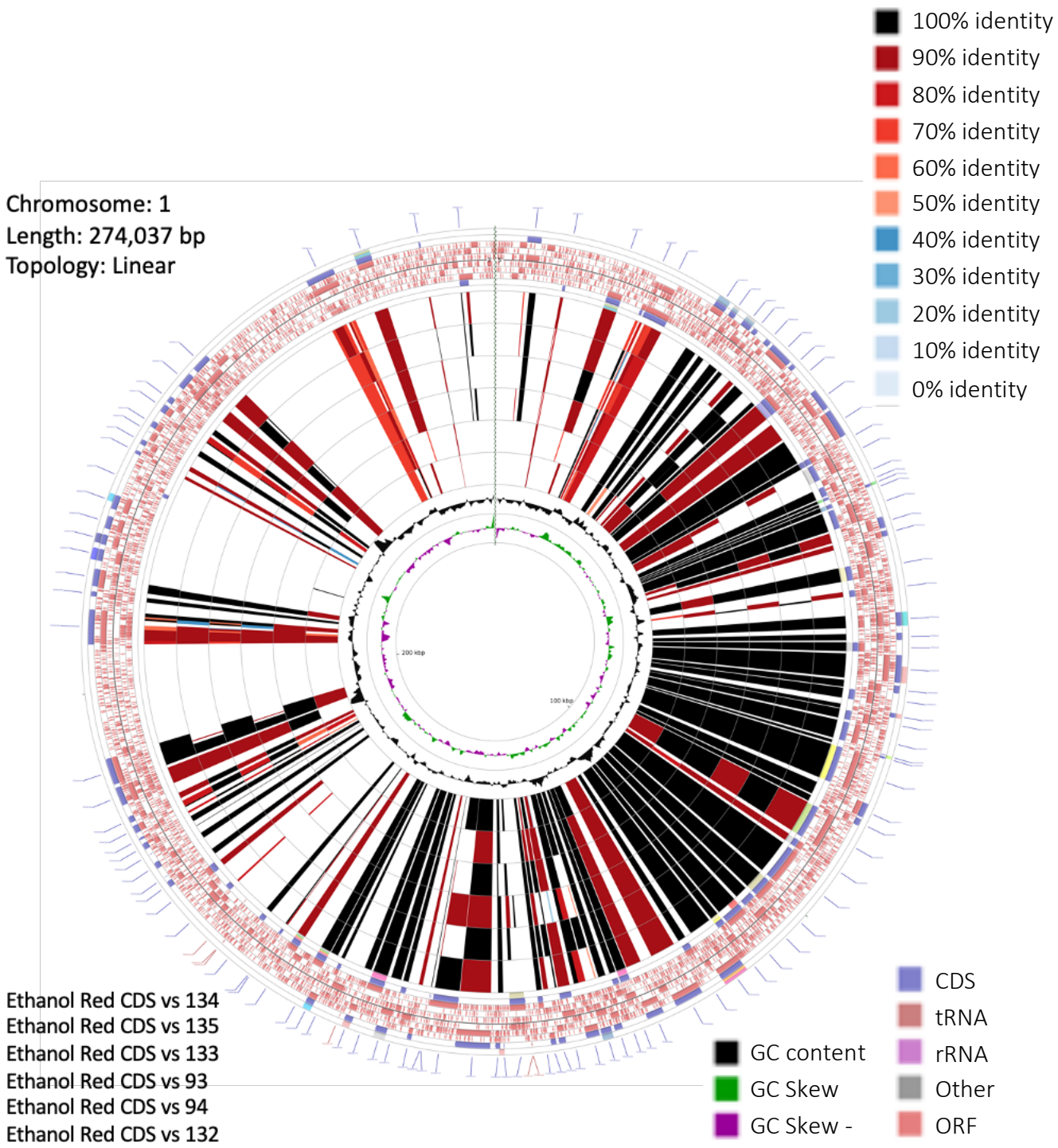


Figure 3.14 CCT map comparing the EtOH Red chromosome 1 with chromosome 1 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring the feature rings depict: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 14. GC content 9. GC skew

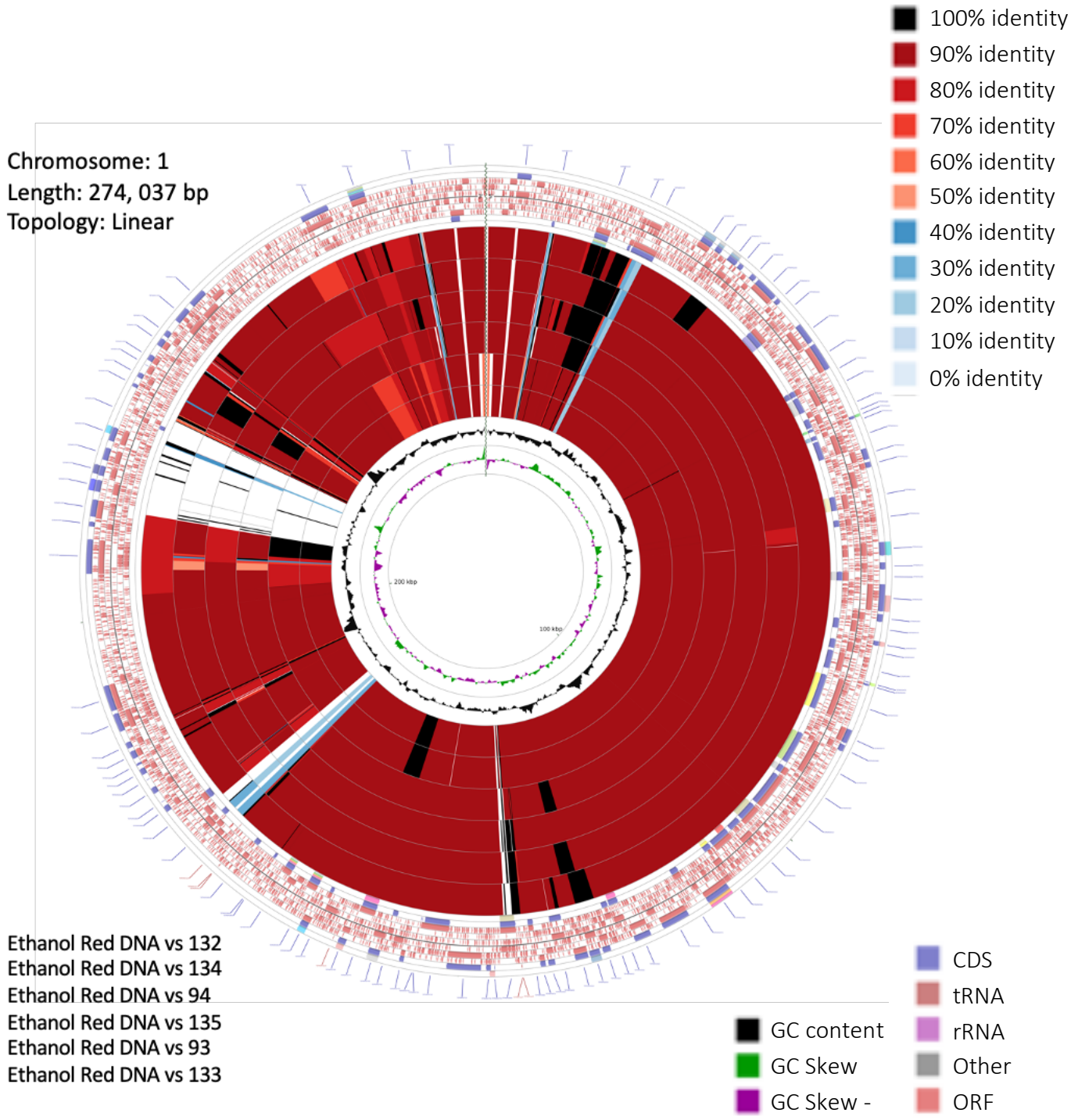


Figure 3.15 CCT map comparing the EtOH Red chromosome 1 with chromosome 1 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring the feature rings depict: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 14. GC content 9. GC skew

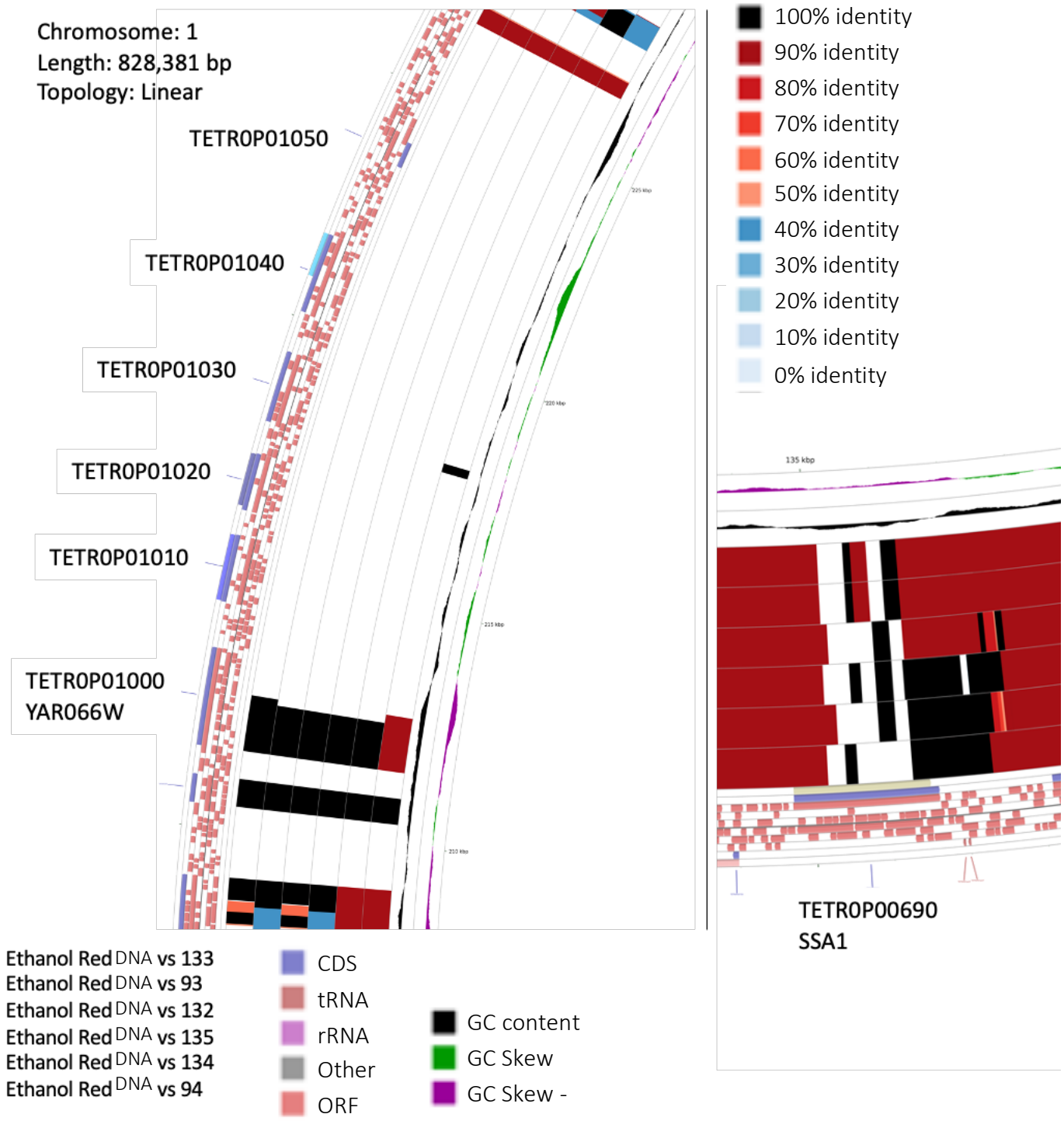


Figure 3.16 Zoomed CCT map of Chromosome 1. Generated using the create_zoomed_maps.sh command from the project used to create figures 3.11 and 3.12. A. is 217,000 bp region zoomed by a factor of 7 and shows. B. is the 135,000 bp region zoomed by a factor of 7.

Chromosome 2 of all six sequenced genomes shows a high level of identity to that of the reference genome with all coding regions showing >60% similarity in at least 2 genomes and above 90% in most predicted coding regions (figure 3.17). As is true throughout this analysis similarity by this comparison will be affected by the truncations indicated by the Ideel analysis discussed previously. This level of similarity is also evident in the DNA vs DNA map (figure 3.18) where only the beginning and end of the chromosomal sequence where there is little to no sequence aligned. The zoomed map in figure 3.19 shows this region and the 4 predicted coding regions that have no similarity in the six sequenced strains and others that have limited similarity through the lineage These were all annotated by YGAP as hypothetical proteins, however, BLASTP search of TETR0G04090 finds that it has 100% identity with 100% query cover to Mal32p which is a maltase gene responsible for the breakdown of maltose into glucose for utilisation (Charron *et al.*, 1989). This is preceded in this assembly by TETR0G04080 which is shown to have <50% nucleotide identity for this coding region in the lineage can be found to have 100% similarity with 100% query cover to MAL61 which is a maltose transport protein. This would suggest that Ethanol Red can utilise maltose and that none of the strains in the lineage can, however when the coverage of Illumina reads mapped to the Ethanol Red reference genome is observed the coverage of reads in this area is significantly lower than in the rest of the chromosome which may be indicative of an assembly error. As such the sequence of Mal32 was searched for in the assembly of strain 94 to determine if this observation is the result of assembly error, 99, 77 and 75% nucleotide identity with 99% query coverage to the Mal32 sequence was found on chromosomes 8, 11 and 6 respectively. Given the known history of these strains it is proposed that this is likely not to be an assembly error but rather the result of genome engineering to improve maltose utilisation by these strains.

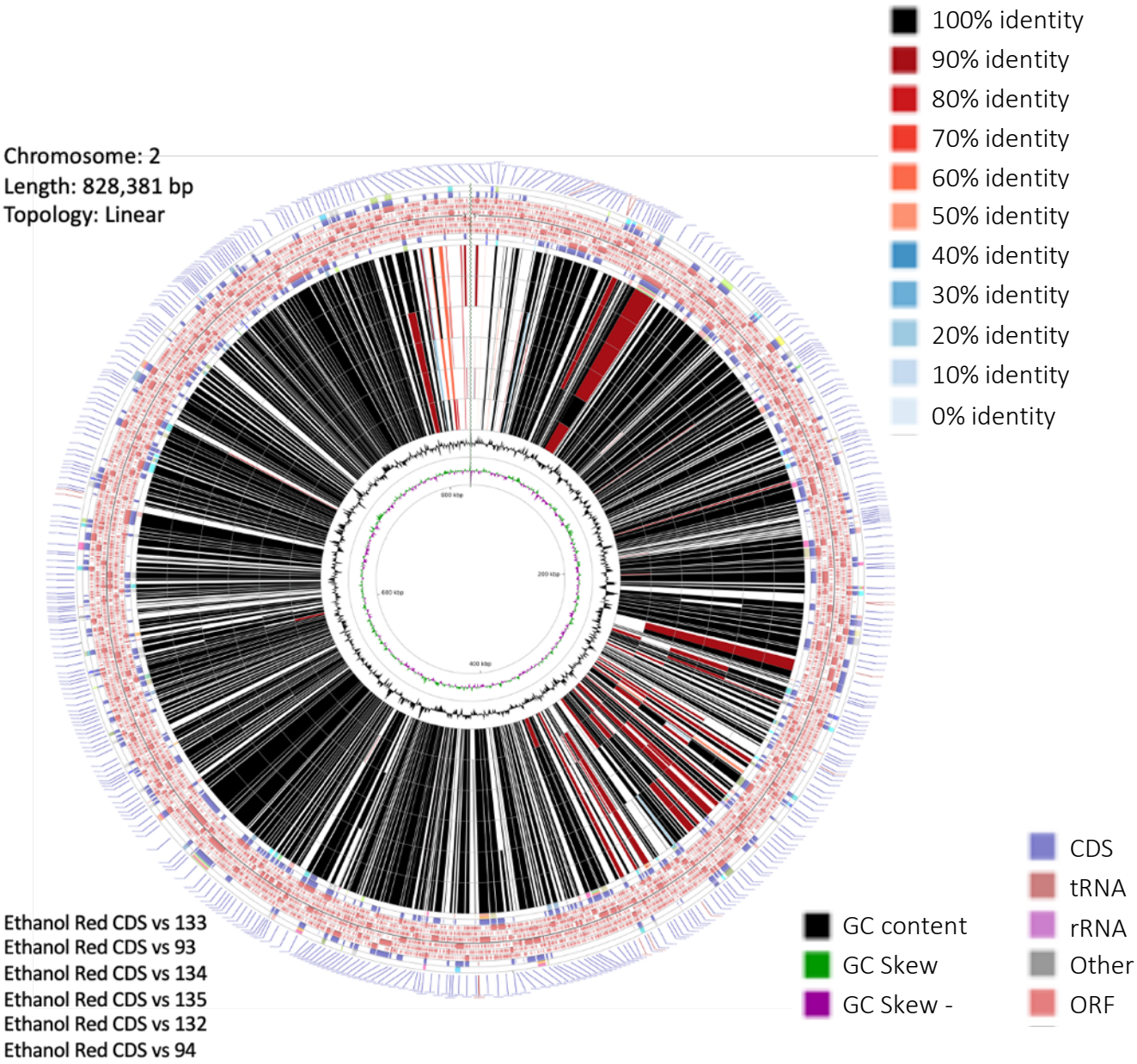


Figure 3.17. CCT map comparing the EtOH Red chromosome 2 with chromosome 2 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring depict: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

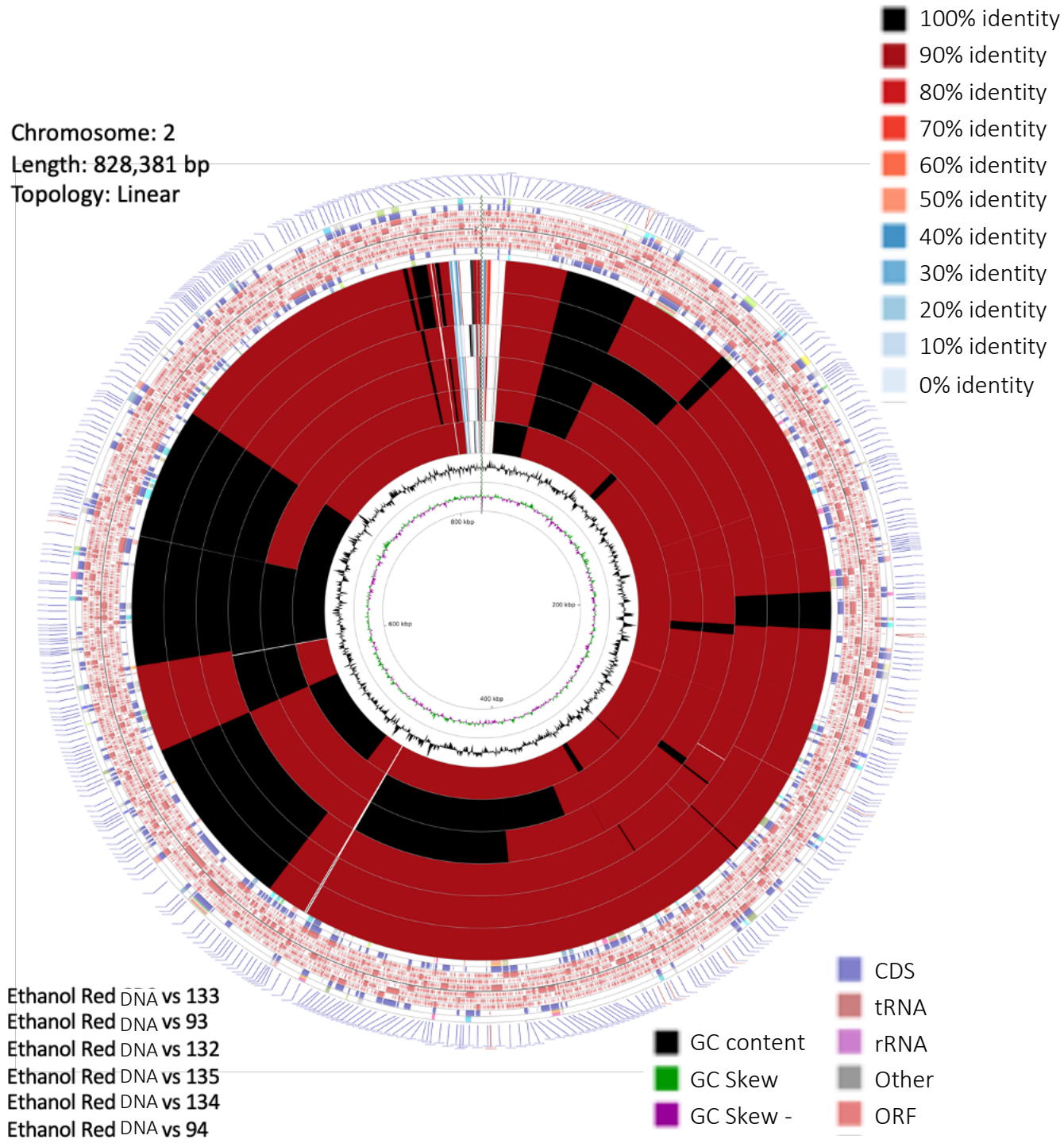


Figure 3.18 CCT map comparing the EtOH Red chromosome 2 with chromosome 2 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring the feature rings depict: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

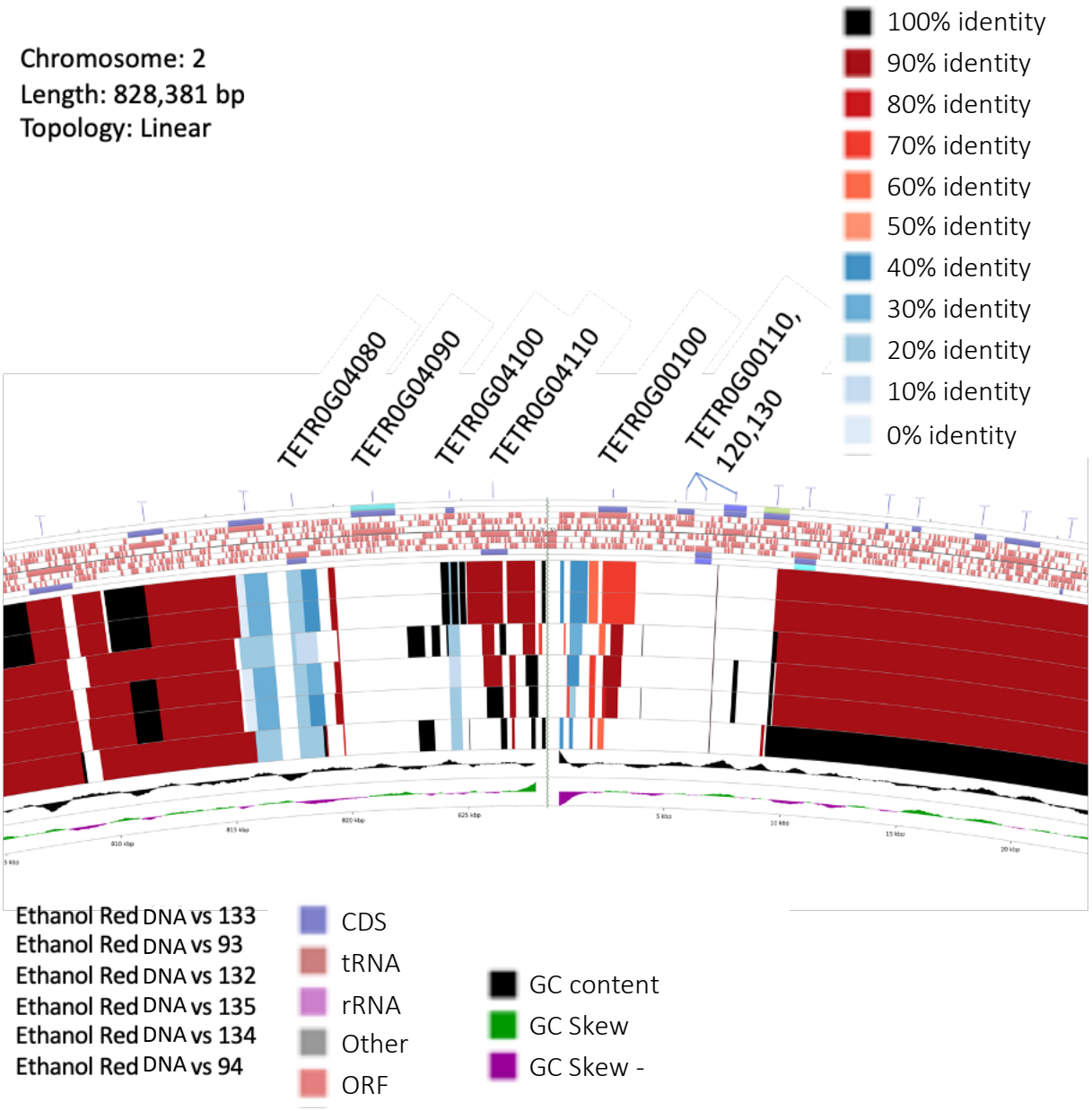


Figure 3.19 Zoomed CCT map of Chromosome 2. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.17 and 3.18. Here is the 0 bp region zoomed by a factor of 9.

Chromosome 3 which is 325 kb in length also has a large level of similarity amongst all strains apart from the beginning and end of the chromosome as well as one distinct region close to the end. The CDS vs CDS map in figure 3.20, shows the high level of similarity between coding regions in the Ethanol Red chromosome 3 and the corresponding chromosome in the six strains of the industrial lineage. The two regions of difference that are spoken about in figure 3.18 are more evident in the DNA vs DNA map given in figure 3.21. These two regions are shown in more detail in the zoomed maps in figure 3.22. The region at 305 kb contains 6 predicted protein-coding sequences that are given the annotations TER0N01600-660 by YGAP. Of these TER0N01600 (YPR202W) is of unknown function but like telomeric helicases. TER0N01630 is annotated with unknown function by YGAP but has 99.28 % identity and 98% query cover to DEAD/DEAH box helicase family protein. The other four are annotated with unknown function by YGAP and have no high similarity matches by BLASTP search. These differences seen are likely due to artefacts of the sequencing and assembly, given that there is a high likelihood that a small section of DNA at the telomeres would have low or no sequencing coverage. There are methods available to improve telomere sequencing that were not employed in this study as the information lost in this way was determined to be minimal, however, it was noted that there is a pattern of under-sampling telomeric sequence with all the lineage strains. An alternative explanation for this could be telomeric rearrangement in yeast which has been reported as happening at a rate of 1 in 20 in *S. cerevisiae* transformants by observing the chromosomal location of telomere encoded genes (Venter and Hörz, 1989).

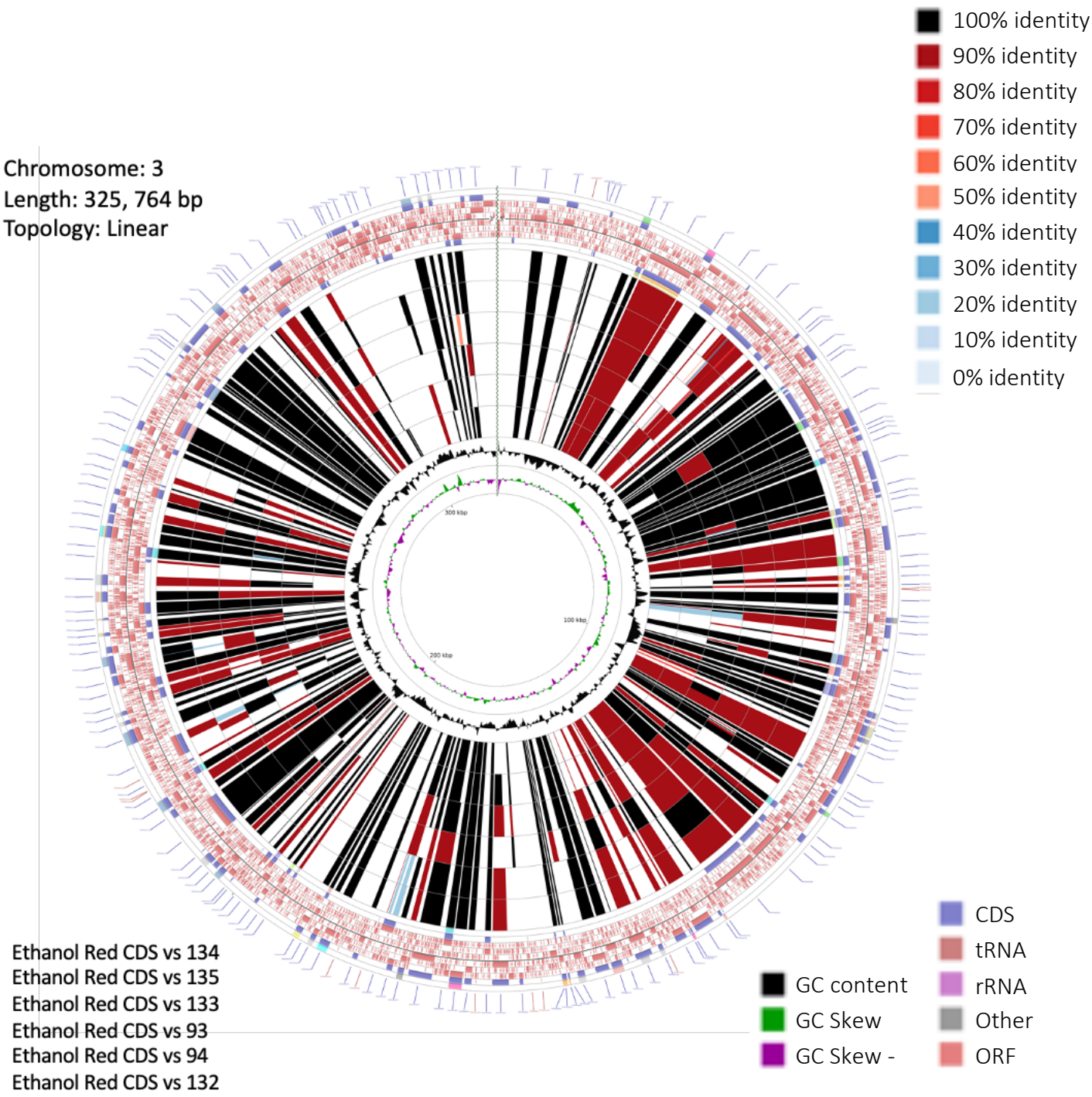


Figure 3.20 CCT map comparing the EtOH Red chromosome 3 with chromosome 3 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost rings depict: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

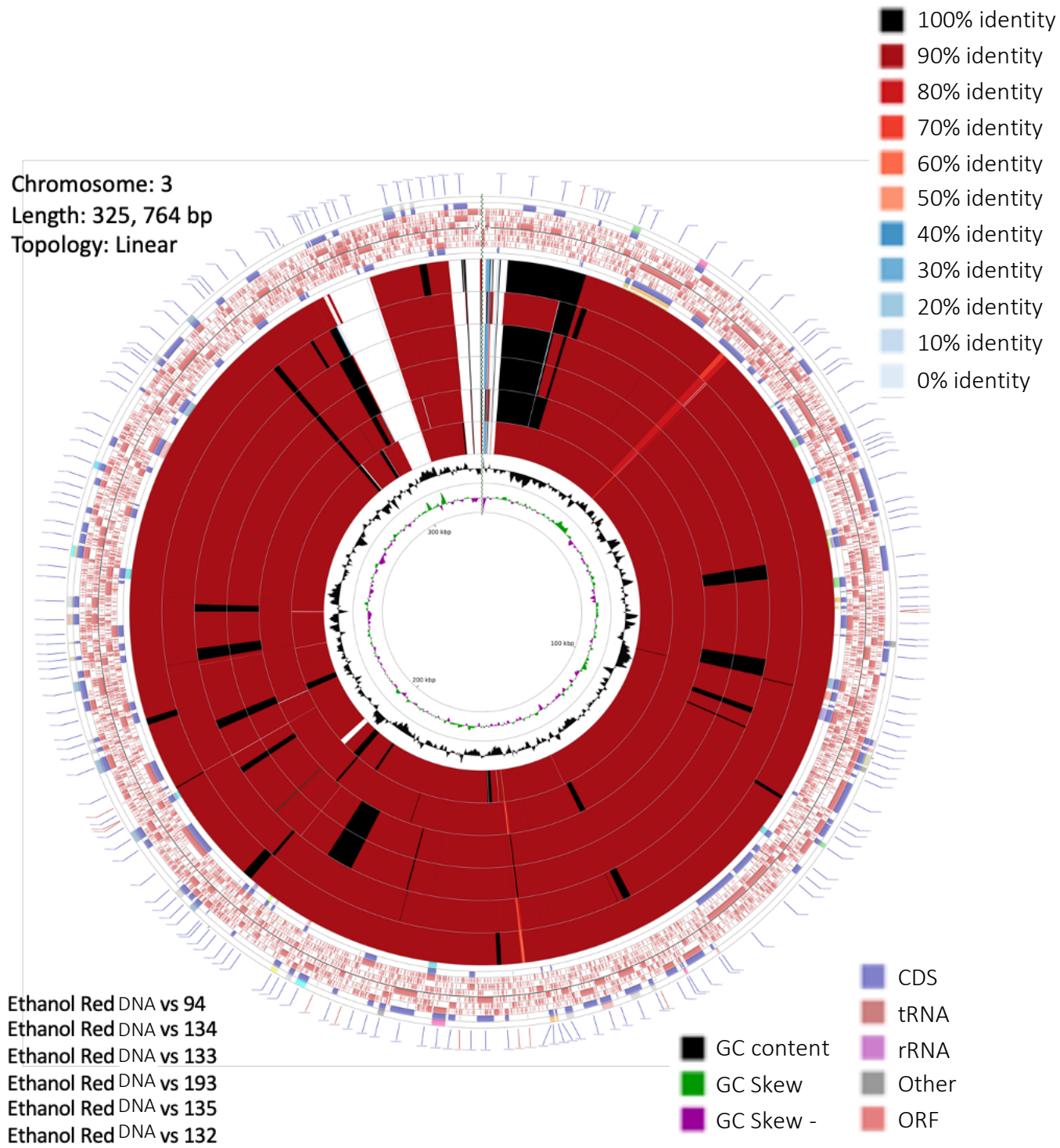


Figure 3.21 CCT map comparing the EtOH Red chromosome 3 with chromosome 3 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

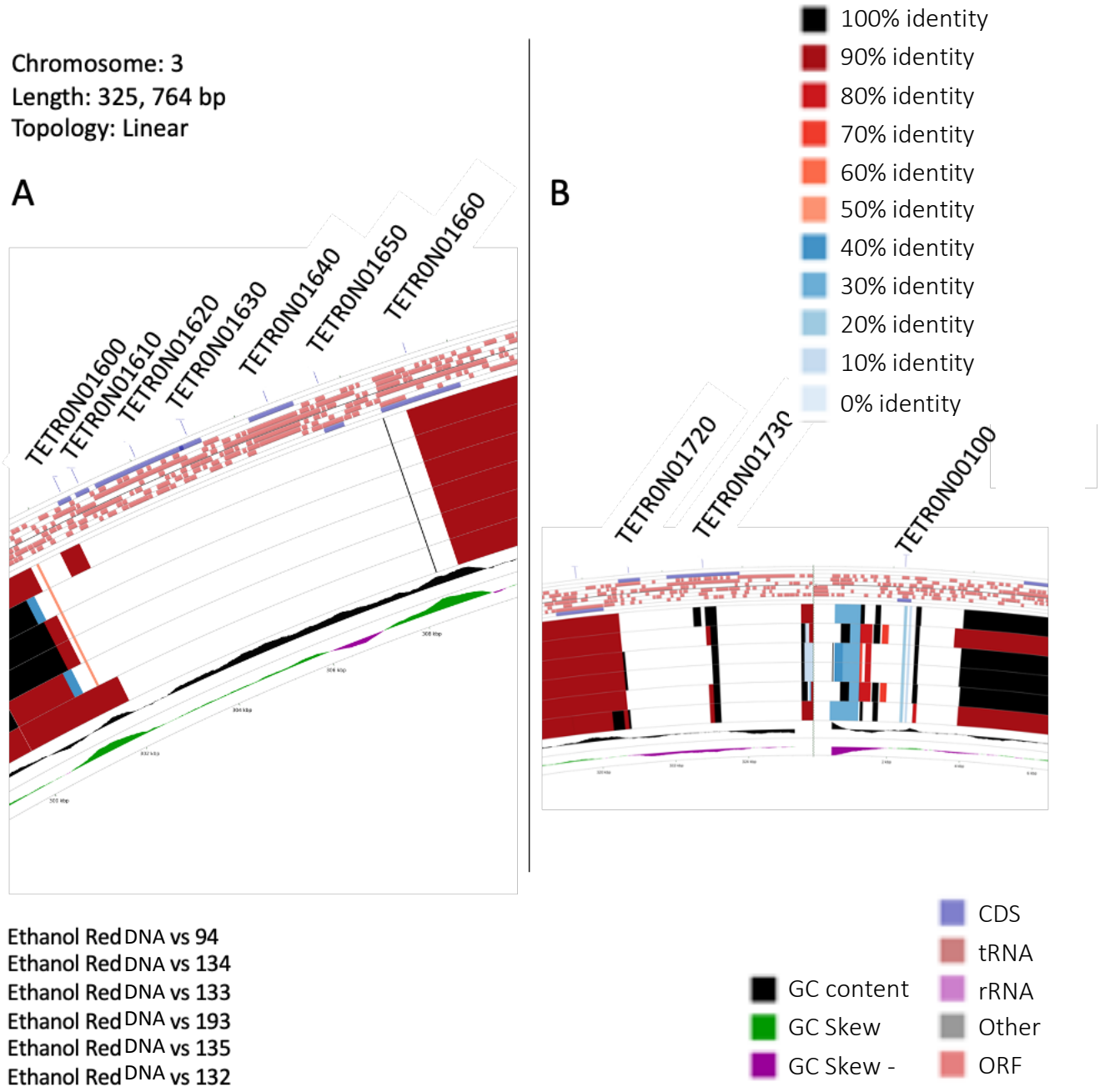


Figure 3.22 Zoomed CCT map of Chromosome 3. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.17 and 3.18. A) is the 305,000 bp region zoomed by a factor of 9. B is the 0 bp region zoomed by a factor of 9.

Chromosome 4, the largest of the *S. cerevisiae* chromosomes at 1,472,829 bp in length still has a very high degree of DNA similarity but as would be expected with a larger number of genes on this chromosome there are also a larger number of protein-coding regions that vary between the seven strains compared here. Figure 3.23 shows the CDS similarity and figure 3.24 shows the DNA vs DNA similarity. The CDS comparison shows that there are several differences in protein coding regions between the strains, this could be explained by errors in base calling in the reference that were not properly corrected by the Illumina reads in Pilon. The Ideel results discussed earlier identified several cases where proteins are truncated or with deletions in these assemblies. There is a good chance that some of these instances are represented in this figure. In the DNA vs DNA map, there are two regions where only one of the strains does not show similarity. In both cases where one strain has a relatively large region of difference, there is a small adjacent region of difference from Ethanol Red that is shared between all six strains of the lineage. This can be seen in more detail in figure 3.22 B. where TETR0A00300 is not found to have homologous sequence in the lineage and adjacent to this is a region of 20 predicted coding regions where all lineage strain apart from 94 show a high degree of identity. Similarly, in Figure 3.25 D. TETR0A05970 has limited DNA identity across all strains and strain 132 is the only one to no/low similarity in the 35 up-stream genes. Figure 3.25 A. and C. give three distinct regions of difference containing a total of 6 predicted coding regions. The predicted gene names and functional information for these regions is given in table 3.6.

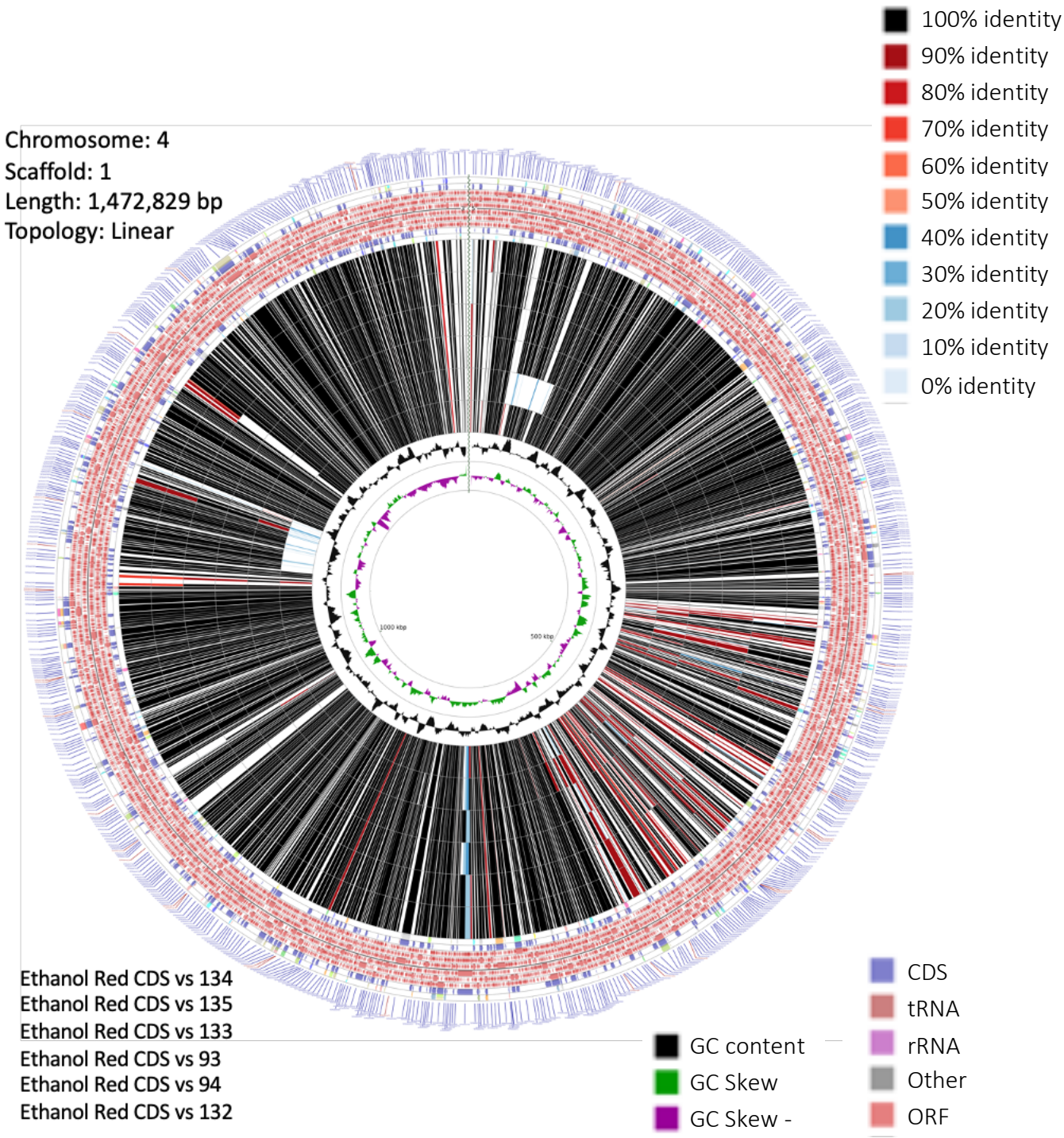


Figure 3.23 CCT map comparing the EtOH Red chromosome 4 with chromosome 4 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 14. GC content 9. GC skew

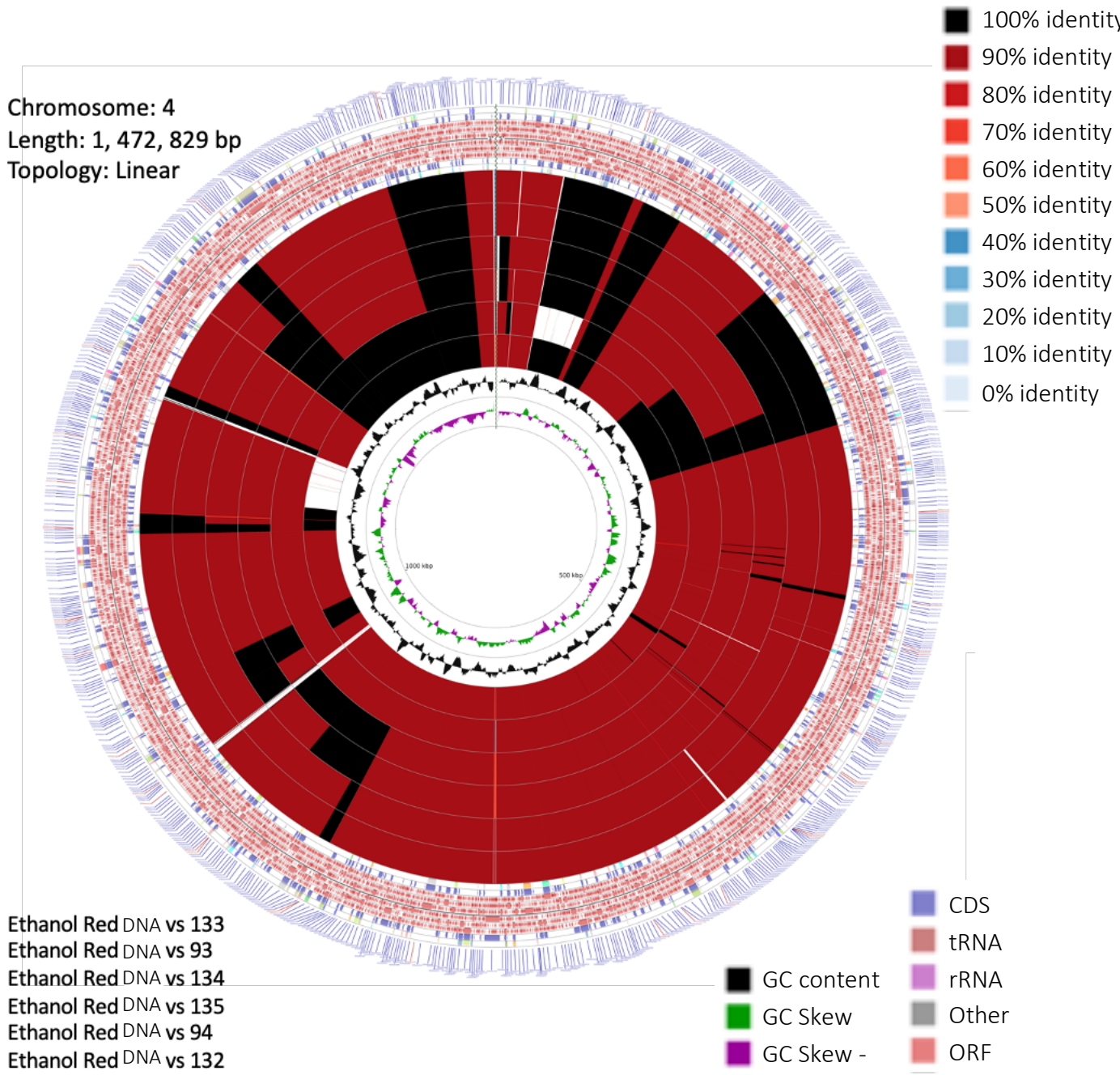


Figure 3.24 CCT map comparing the EtOH Red chromosome 4 with chromosome 4 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 14. GC content 9. GC skew

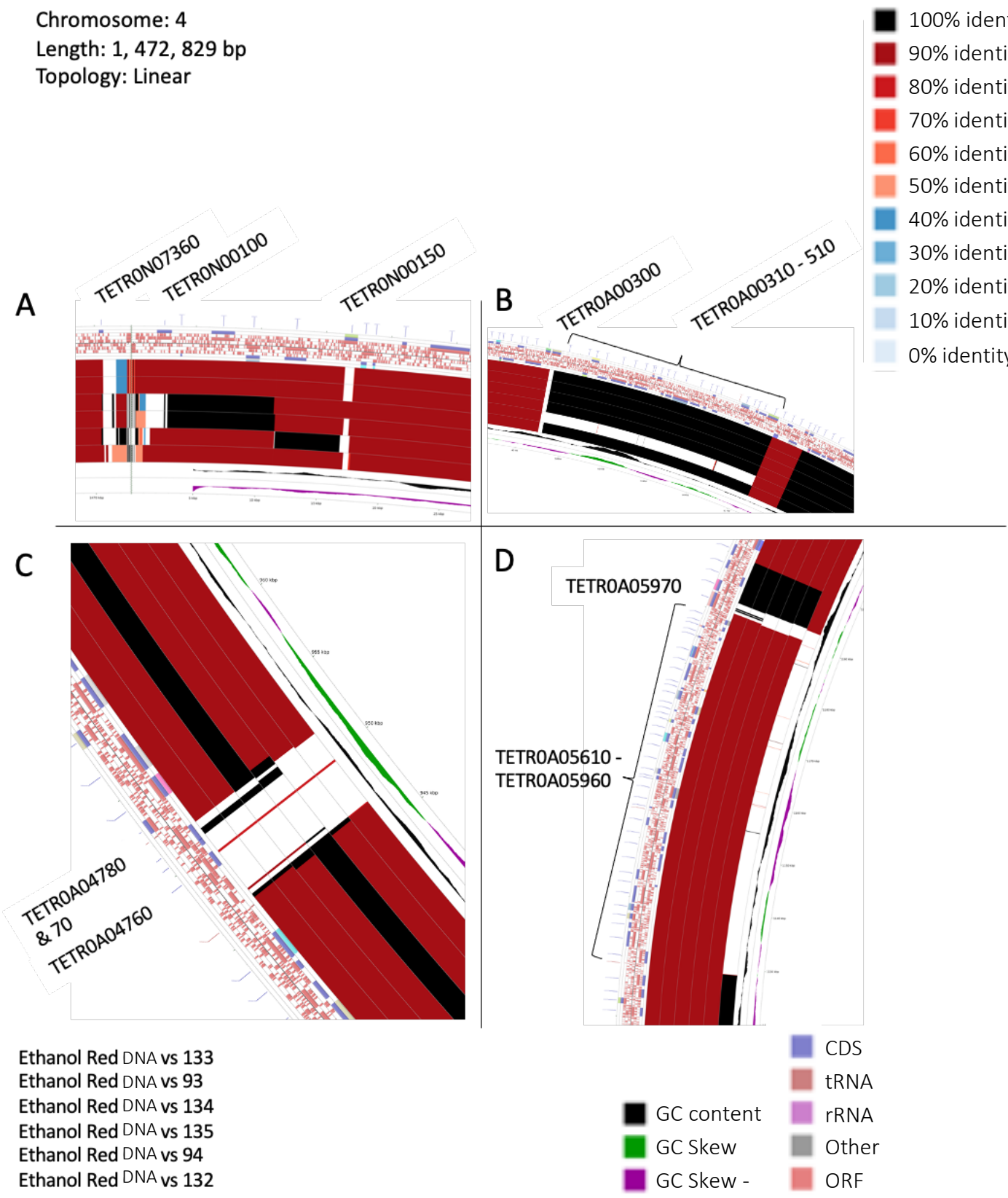


Figure 3.25 Zoomed CCT map of Chromosome 3. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.20 and 3.21. A) is the 18,000 bp region zoomed by a factor of15. B) is the 70000 bp region zoomed by a factor of 9. C) is the 950,000 bp region zoomed by a factor of 15. D) is the 1,300,000 bp region zoomed by a factor of 9.

YGAP Annoation	Predicted Gene	Standard Name	Function	Reference
TETR0A07360	YDR542W	PAU10	seriPAUperin	(Viswanathan et al. 1994)
TETR0A00100	Hypothetical Protein			
TETR0A00150	Hypothetical Protein			
TETR0A00300	DL229W,YNL209W	SSB2	Stress-Seventy subfamily B	(Nelson et al. 1992)
TETR0A00310	YDL227C	HO	Homothallic switching endonuclease	(Jensen, Sprague, and Herskowitz 1983)
TETR0A00320	YNL204C	SPS18	SPorulation-Specific	(Coe, Murray, and Dawes 1994)
TETR0A00330	YDL225W	SHS1	Seventh Homolog of Septin	(Mino et al. 1998)
TETR0A00340	YDL224C,YNL197C	WHI3	Cell cycle regulation, RNA binding protein	(Nash, Volpe, and Futcher 2001)
TETR0A00350	YNL195C,YDL223C	HBT1	HuB1 Target	(Dittmar et al. 2002)
TETR0A00360	YDL222C,YNL194C		Membrane protein required for sporulation	(Byrne and Wolfe 2005b)
TETR0A00370	YDL220C	CDC13	Cell Division Cycle, ssDNA binding protein, telomerase regulation.	(Hartwell, Culotti, and Reid 1970)
TETR0A00380	YDL219W	DTD1	D-Tyr-tRNA(Tyr) Deacylase	(Soutourina, Blanquet, and Plateau 2000)
TETR0A00390	YDL218W	-	Unknown, induced by starvation	
TETR0A00400	YDL217C	TIM22	Translocase of the Inner Mitochondrial membrane	(Sirrenberg et al. 1996)
TETR0A00410	YDL216C	RRI1	Regulator of Rub1 specific Isopeptidase	(Lyapina et al. 2001)
TETR0A00420	YDL215C	GDH2	Glutamate DeHydrogenase	(Miller and Magasanik 1990)
TETR0A00430	YDL214C,YNL183C	NPR1	Nitrogen Permease Reactivator	(VANDENBOL et al. 1987)
TETR0A00440	YDL213C	NOP6	NucleOlar Protein	(Samanta and Liang 2003)
TETR0A00450	YDL212W	SHR3	Super high Histidine Resistant	(Ljungdahl et al. 1992)
TETR0A00460	YDL211C	-	Unknown	
TETR0A00470	YDL210W	UGA4	Utilization of GAba	(RAMOS et al. 1985)
TETR0A00480	YDL209C	CWC2	Complexed With Cef1p	(Ohi et al. 2002)
TETR0A00490	YDL208W	NHP2	Non-Histone Protein	(Kolodrubetz and Burgum 1991)
TETR0A00500	YDL207W	GLE1	GLFG (glycine-leucine-phenylalanine-glycine) LEthal	(Murphy, Watkins, and Wente 1996)
TETR0A00510	YDL206W		Unknown	
TETR0A04760	Hypothetical Protein		Ty like retrotransposon	
TETR0A04770	Hypothetical Protein		Ty like retrotransposon	
TETR0A04780	Hypothetical Protein			
TETR0A05610	YDR350C	ATP22	ATPase synthase	(Helfenbein et al. 2003)
TETR0A05620	YDR351W,YHR103W	SBE22	similar to SBE2, transport of cell wall components	(Santos and Snyder 2000)
TETR0A05630	YDR352W	YPQ2	Yeast PQ-loop protein, nitrogen stress	(Ježégou et al. 2012)
TETR0A05640	YDR353W,YHR106W	TRR2	ThioRedoxin Reductase	(Machado, Morgan, and Merrill 1997)
TETR0A05650	YDR354W	TRP4	TRyPtophan, tryptophan biosynthesis	(Miozzari, Niederberger, and Huetter 1978)
TETR0A05660	YDR356W	SPC110	Spindle Pole Component	(Kilmartin et al. 1993)
TETR0A05670	YDR357C	CNL1	CNo-Like	(Hayes et al. 2011)
TETR0A05680	YDR358W,YHR108W	GGA1	GGA2, Golgi transport	(Byrne and Wolfe 2005b)
TETR0A05690	YDR359C	EAF1	Esa1p-Associated Factor	(Nourani et al. 2001)
TETR0A05700	YDR361C	BCP1	cytoskeleton organization and vesicular transport	(Audhya and Emr 2003)
TETR0A05710	YDR362C	TFC6	Transcription Factor C	(Swanson et al. 1991)
TETR0A05720	YDR363W	ESC2	Establishment of Silent Chromatin	(Dhillon and Kamakaka 2000)
TETR0A05730	YDR363W-A	SEM1	Suppressor of Exocyst Mutations	(Jäntti et al. 1999)ydr
TETR0A05740	YDR364C	CDC40	Cell Division Cycle	(Kassir and Simchen 1978)
TETR0A05750	YDR365C	ESF1	Eighteen S rRNA Factor, pre-rRNA processing	(Peng et al. 2004)
TETR0A05760	YDR367W	KEI1	Kex2-cleavable protein Essential for Inositol phosphorylceramide synthesis	(Sato, Noda, and Yoda 2009)12
TETR0A05770	DR368W,YOR120W	GCY1	Galactose-inducible Crystallin-like Yeast protein	(Angermayr and Bandlow 2003)
TETR0A05780	YDR369C	XRS2	X-Ray Sensitive, cell cycle regulation	(Ivanov, Korolev, and Fabre 1992)
TETR0A05790	YDR370C	DXO1	Decapping eXOnuclease	(Chang et al. 2012)
TETR0A05800	YDR371W	CTS2	ChiTinaSe	(Dünkler et al. 2005)
TETR0A05810	YDR372C	VPS74	Vacuolar Protein Sorting	(Bonangelino, Chavez, and Bonifacino 2002)
TETR0A05820	YDR373W	FRQ1	FReQuenin homolog	(Hendricks et al. 1999)
TETR0A05830	YDR374C	PHO92	PHOsphate metabolism	(Kang et al. 2014)
TETR0A05840	YDR374W-A	WIP1	W-like Protein	(Cliften et al. 2003)
TETR0A05850	YDR375C	BCS1	ubiquinol-cytochrome c reductase (bc1) Synthesis, Complex III assembly	(Nobrega, Nobrega, and Tzagoloff 1992)
TETR0A05860	YDR376W	ARH1	Adrenodoxin Reductase Homolog, iron homeostasis	(Manzella, Barros, and Nobrega 1998)
TETR0A05870	YDR377W	ATP17	ATP synthase	(Arnold et al. 1999)
TETR0A05880	YDR378C	LSM6	Like SM, mRNA decay	(Mayes et al. 1999)
TETR0A05890	YDR379W,YOR127W	RGA1	Rho GTPase Activating Protein, bud site selection	(Stevenson et al. 1995)
TETR0A05900	YDR379C-A	SDH6	succinate dehydrogenase assembly	(Na et al. 2014)
TETR0A05910	YDR380W	ARO10	AROmatic amino acid requiring, Phenylpyruvate decarboxylase	(Iraqui et al. 1999)
TETR0A05920	YDR381W	YRA1	Yeast RNA Annealing protein	(Portman, O'Connor, and Dreyfuss 1997)
TETR0A05930	YDR381C-A	COI1	Cytochrome c Oxidase Interacting protein, Mitochondrial assembly factor	(Singhal et al. 2017)
TETR0A05940	YDR382W	RPP2B	Ribosomal Protein P2 Beta	(Planta and Mager 1998)
TETR0A05950	YDR383C	NKP1	Non-essential Kinetochore Protein	(Cheeseman et al. 2002)
TETR0A05960	YDR384C	ATO3	Ammonia (Ammonium) Transport Outward	(Palková et al. 2002)
TETR0A05970	YDR385W,YOR133W	EFT1	Elongation Factor Two, ribosomal translocation	(Byrne and Wolfe 2005b)

Table 3.8. Annotation information for predicted coding regions in chromosome 4

Chromosome 5, 568, 786 bp, for which the CDS vs CDS map is given in figure 3.26 and the DNA vs DNA is given in figure 3.27 there are no large regions of differentiation. The five small regions where there is no or little similarity between the lineage and the reference Ethanol Red are shown in figure 3.28. A shows TETR0K00130 which is YEL070W (DSF1), a Mannitol dehydrogenase (Ohkuni *et al.*, 2006) beside TETR0K00140 (YEL069C/HXT13) a protein involved in the uptake of mannitol (Kruckeberg, 1996). In this region is also TETR0K00150 which is. annotated as unknown by YGAP but has 100 % identity and 100% query cover to Hpa3p, an N-acetyltransferase (Vijayraghavan, Company and Abelson, 1989). These three genes are absent in all 6 of the comparison strains and would suggest that these lack or have reduced ability to utilise mannitol. B. shows a region where all or part of two genes are absent in the lineage when compared to Ethanol Red. This includes, TETR0K00520 which has 99 % identity and 98% query cover to a gag-pol fusion protein and TETR0K00530 with 98.7% identity and 92% query cover to retrotransposon YGR161C-D. figure 3.28 D. shows a region containing seven genes that are found to have high nucleic acid similarity to strains 93, 94, 132 and 133 but no identity to 134 and 135. The genes included in this regions are as follows;TETR0K00920 (YER012W/PRE1) the beta 4 subunit of the 20S proteasome (Heinemeyer *et al.*, 1991); TETR0K00930 (YER013W/PRP22) a DEAH-box RNA-dependent ATPase; TETR0K00940 (YER014W;HEM14) which is involved in heme-biosynthetesis (Urban-Grimal and Labbe-Bois, 1981); TETR0K00950 (YER015W/FAA2) a fatty acyl-CoA synthetase (Johnson *et al.*, 1994); TETR0K00960(YER016W/BIM1) a microtubual binding and regulation (Schwartz, Richards and Botstein, 1997); TETR0K00970 (YER017C;AFG3) which is involved with the degradation of misfolded proteins (Guelin, Rep and Grivell, 1994) and TETR0K00980 (YER018C/SPC25) involved in chromosome segregation (Wigge *et al.*, 1998). Finally figure 3.28 D. Shows a region of high identity with all strains in the lineage flanked by two absent genes; TETR0K02180 and TETR0K02220 which show 100 % identity, 97% query cover and 100 % identity,99.87% query cover respectively to gag-pol fusion protein. All these regions mentioned were found at least once in other locations on the same chromosome by blast search, suggesting again some assembly error.

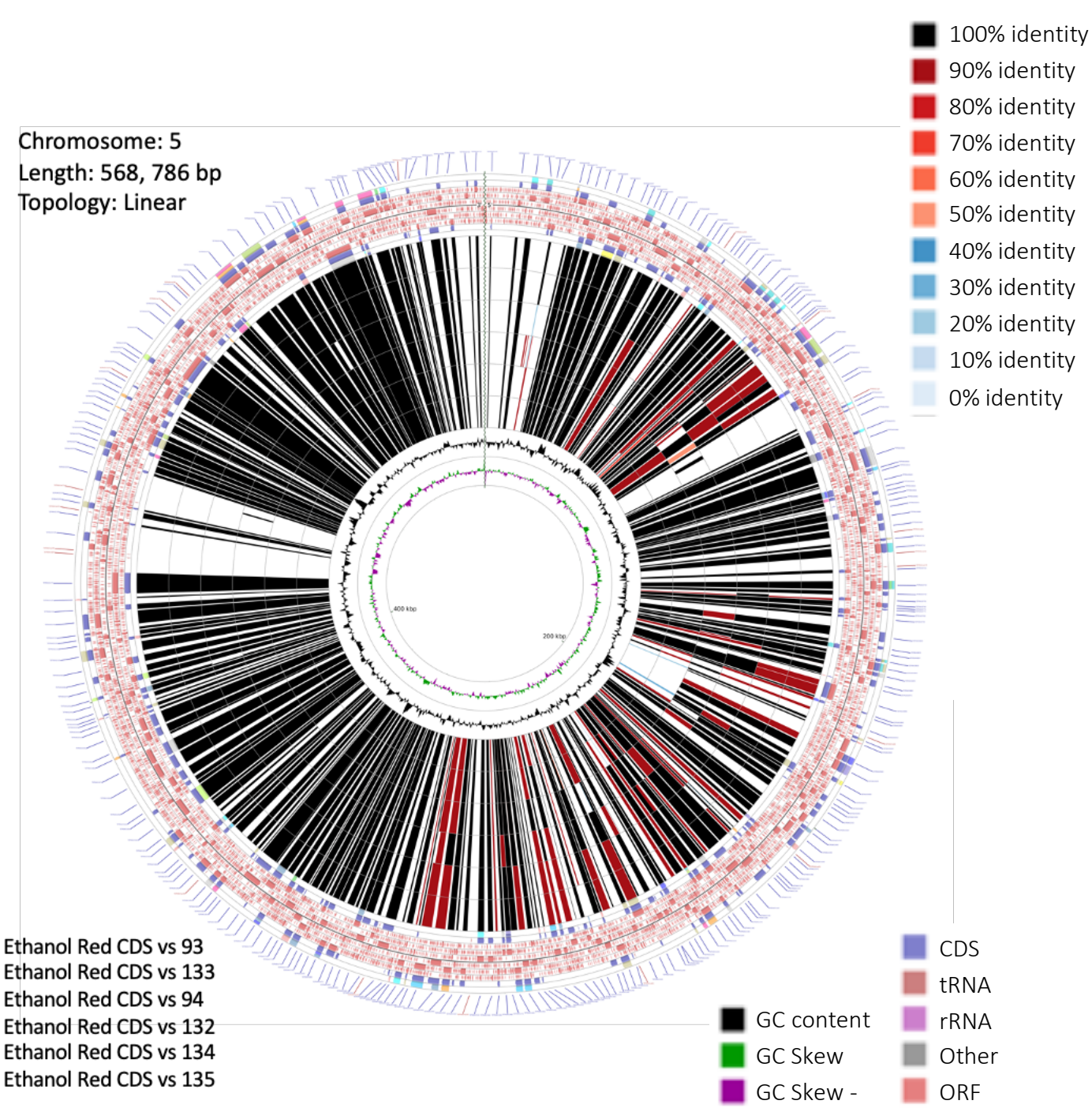


Figure 3.26 CCT map comparing the EtOH Red chromosome 5 with chromosome 5 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

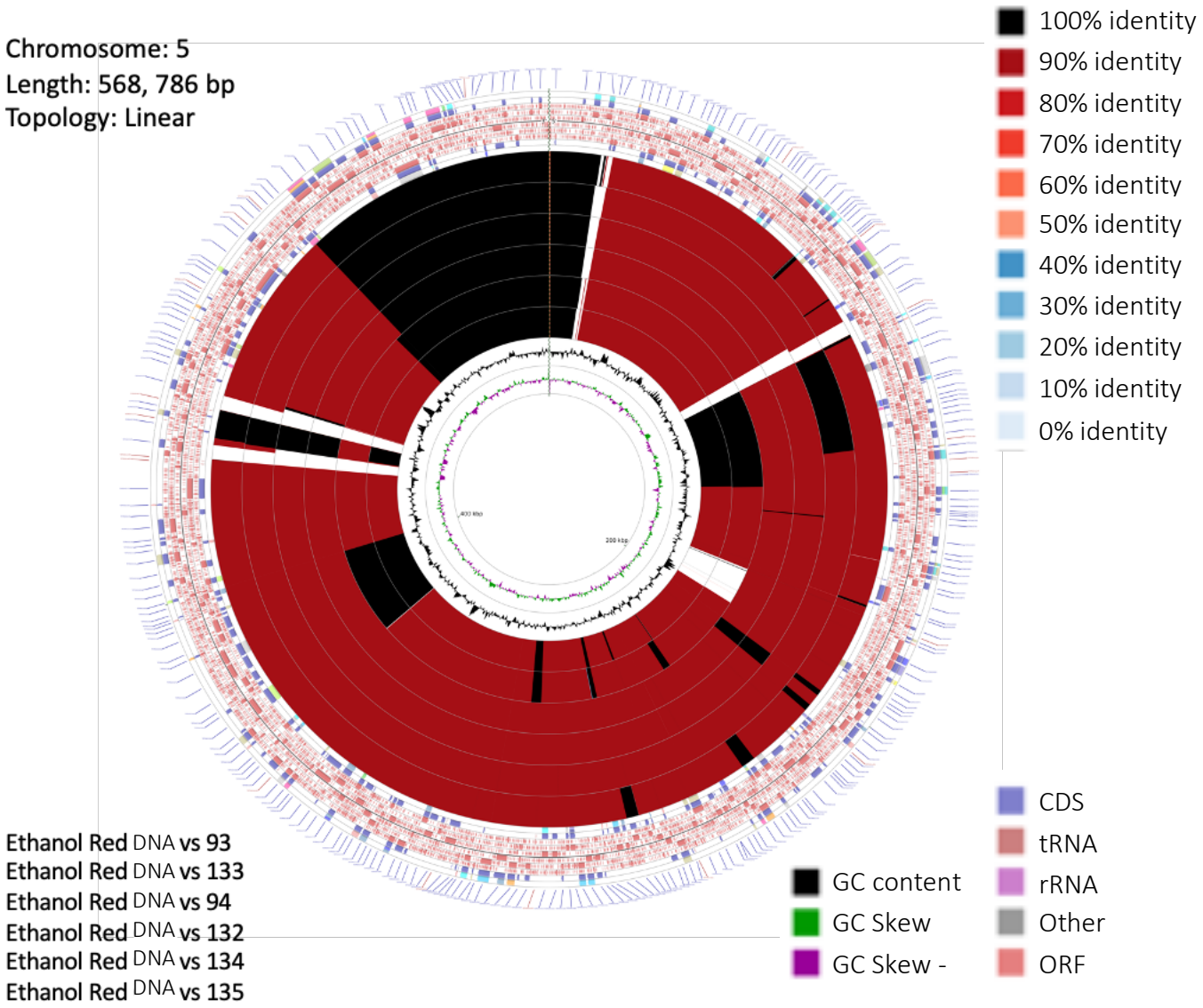


Figure 3.27 CCT map comparing the EtOH Red chromosome 5 with chromosome 5 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

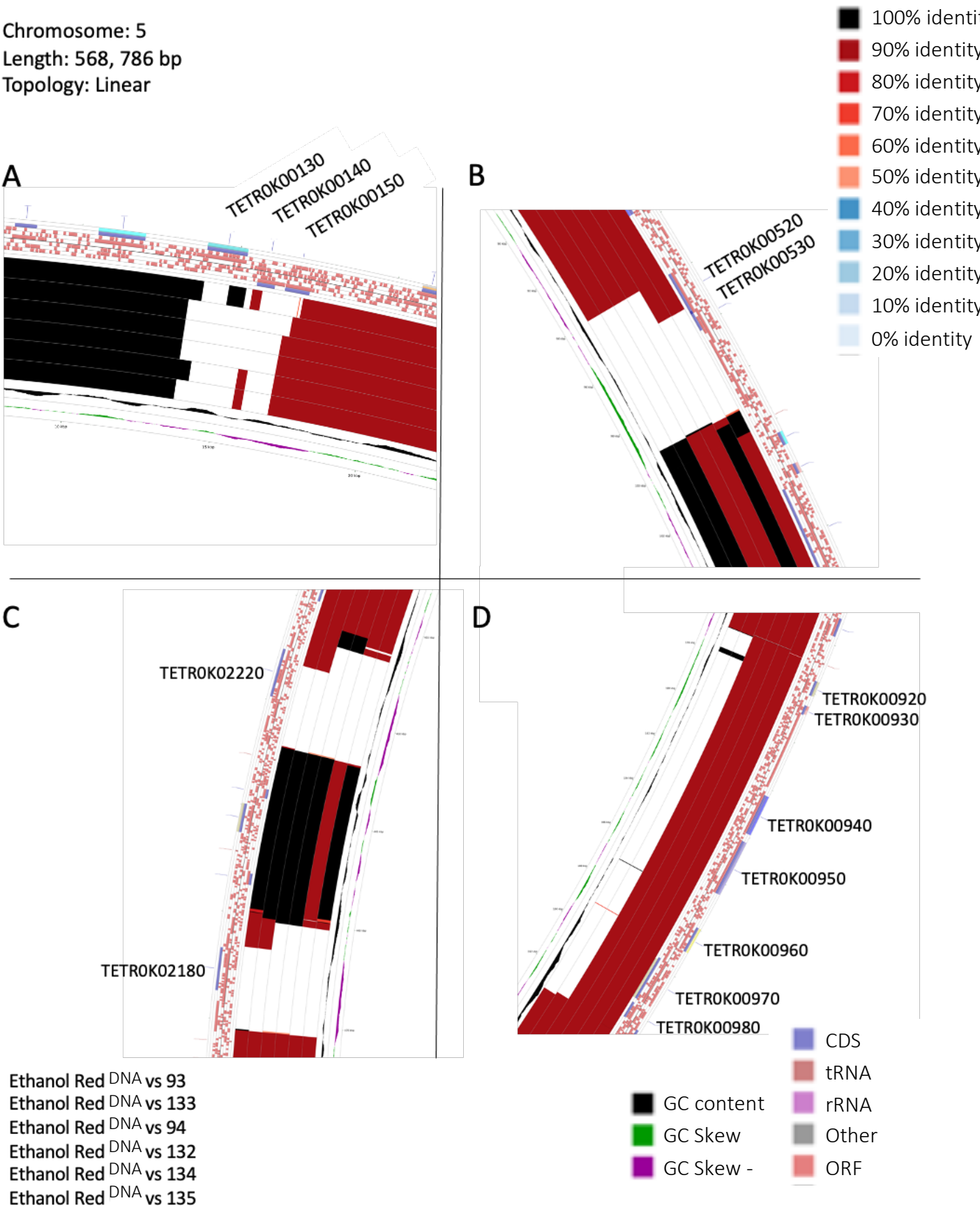


Figure 3.28 Zoomed CCT map of Chromosome 5. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.23 and 3.24. A) is the 10,000 bp region zoomed by a factor of 15. B) is the 75,000 bp region zoomed by a factor of 9. C) is the 190,000 bp region zoomed by a factor of 15. D) is the 440,000 bp region zoomed by a factor of 9.

Chromosome 6,309,050 bp in length has only one region where the lineage is not high in similarity to the reference. This is not obvious immediately from figure 3.29 where the CDS vs CDS map is shown. With this being a smaller chromosome, the predicted coding regions are less dense on the map giving the appearance of less identity. The DNA vs DNA map however, given in figure 3.30, shows a consistently high DNA level identity between all strains, except for one region. This is from the very beginning of the Ethanol Red sequence to around 32 kb and it includes 11 predicted coding regions. Shown in figure 3.31 TETR0O00100 which shows 99.85 % identity, 93% query cover to DEAD/DEAH box helicase family protein. TETR0O00110 which shows 100 % identity, 100% query cover to Pau14p(Luo and van Vuuren, 2009). TETR0O00120 which shows 99.72 % identity, 99% query cover to signal sequence-binding protein or Vth2p a putative membrane binding protein (Cooper and Stevens, 1996). TETR0O00130 which shows 100 % identity, 98% query cover to Ima2p an isomaltase (Teste, Marie François and Parrou, 2010). TETR0O00140 which shows 89.30 % identity, 98% query cover to hexose transporter LGT3 (HXT11 / YOL156W). TETR0O00150 which shows 100 % identity, 100% query cover to another hexose transporter HXT9 (YJL219W). TETR0O00160 which shows 97.6% identity, 100% query cover to putative acetyltransferase, YJL218W. TETR0O00170 which shows 99.49% identity, 100% query cover to REE1 (YJL217W) which regulates the expression of phosphopyruvate hydratase. TETR0O00180 is YJL216C (IMA5)an alpha-glucosidase (Teste, Marie François and Parrou, 2010). TETR0O00190 is HXT8 (YJL214W), a putative hexose transporter (Kruckeberg, 1996). TETR0O00200 which shows 99.41% identity, 100% query cover to YNL332W (THI12) which is involved in thiamine biosynthesis. Again, BLAST searching of these sequences has found them to be present elsewhere on the same chromosome and others in some cases with high (>99%) levels of nucleotide identity. This is probably indicative of misassembly.

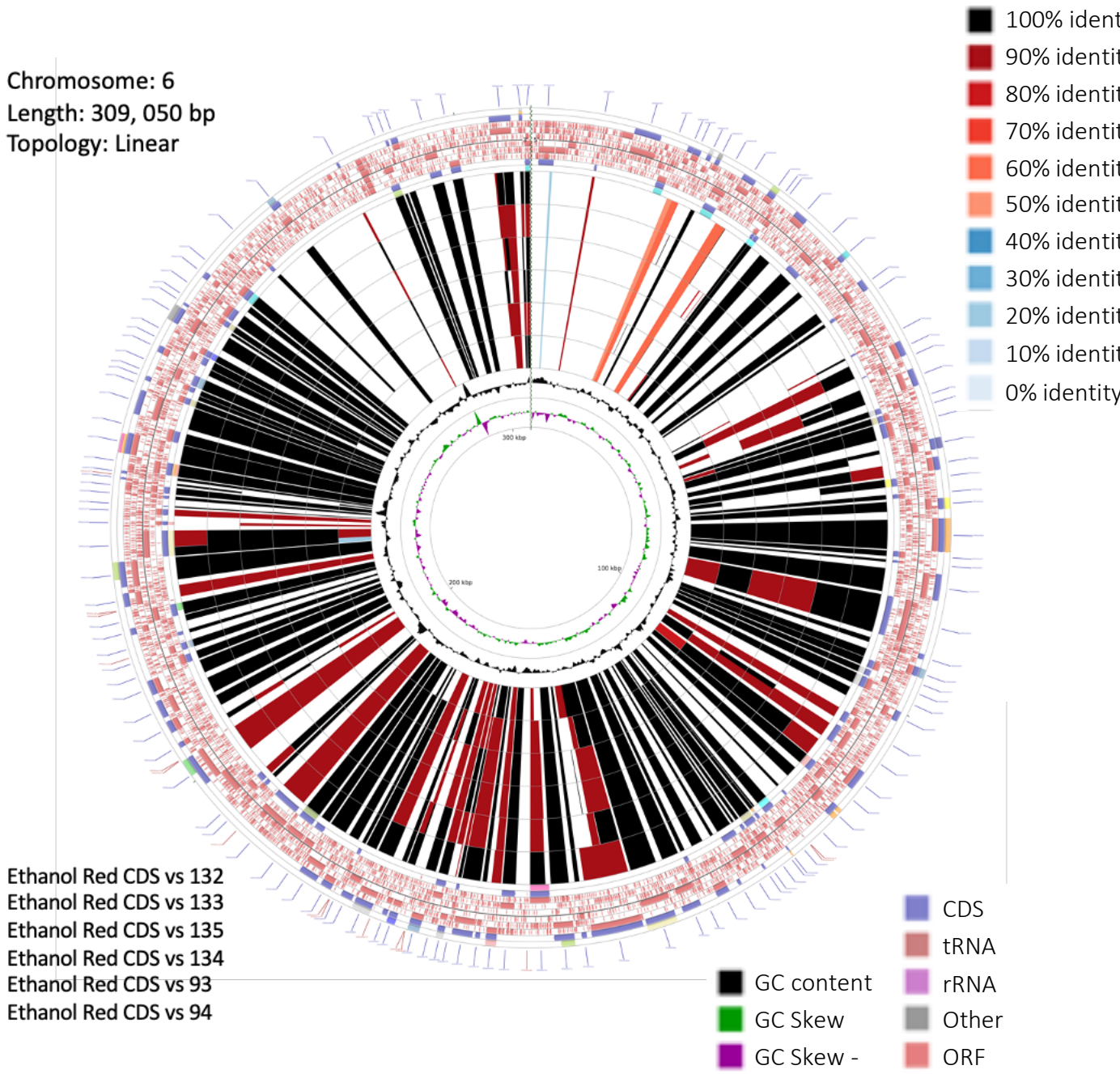


Figure 3.29 CCT map comparing the EtOH Red chromosome 6 with chromosome 6 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

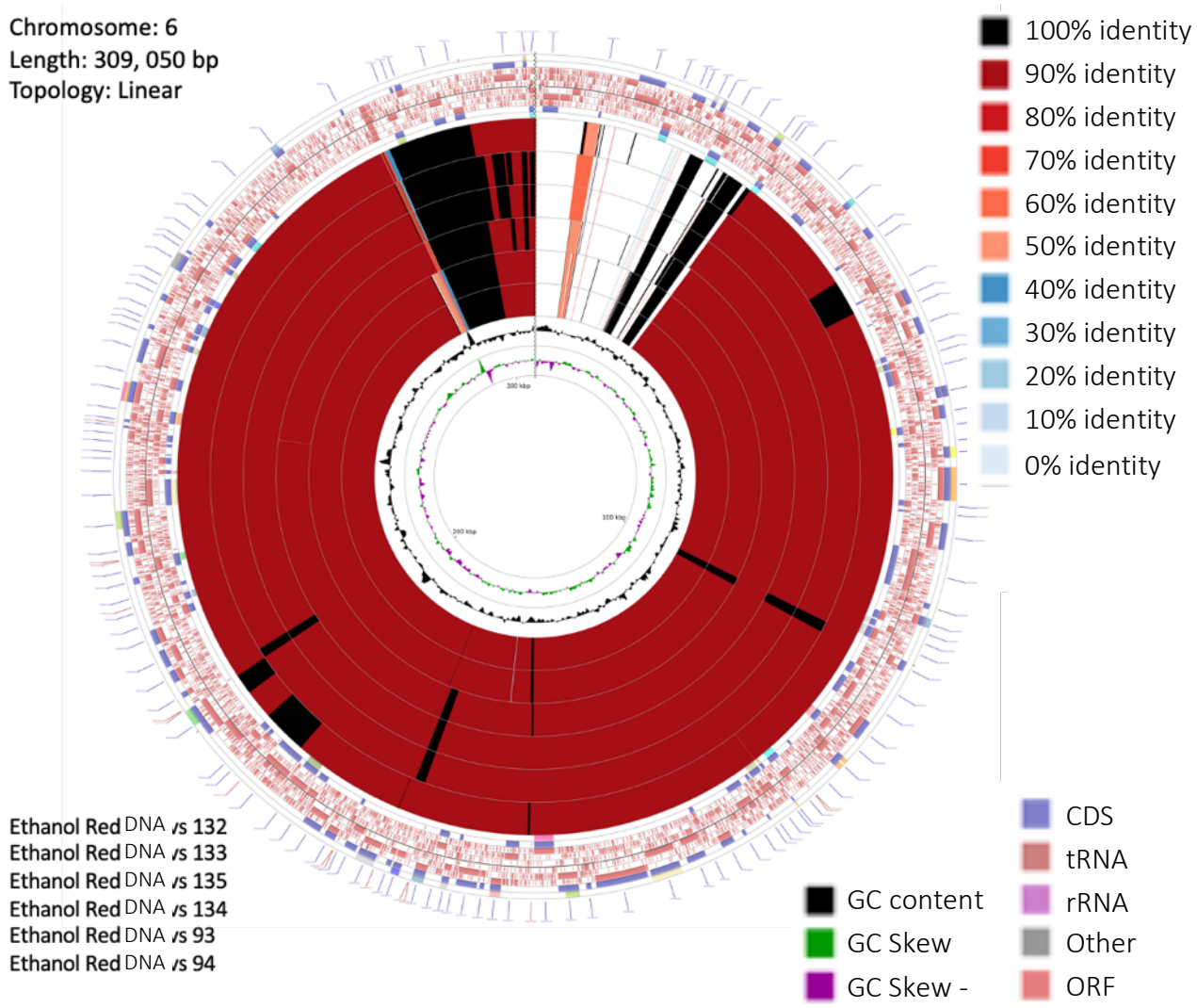


Figure 3.30 CCT map comparing the EtOH Red chromosome 6 with chromosome 6 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

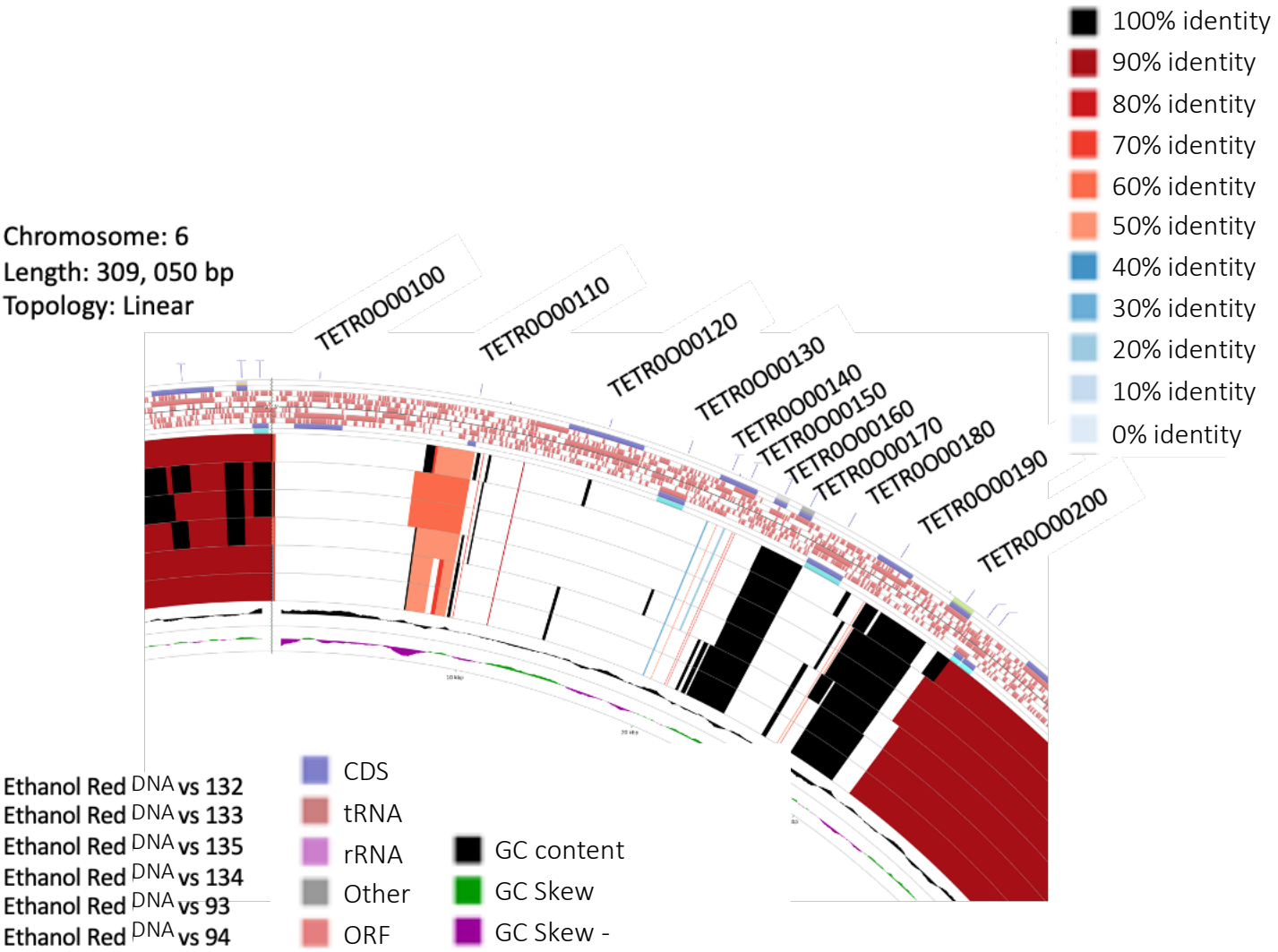


Figure 3.30 Zoomed CCT map of Chromosome 6. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.25 and 3.26. A) is the 18,000 bp region zoomed by a factor of 15.

Chromosome 7, the third largest Ethanol Red chromosome at 1,065,431 bp, shown in figure 3.32 as a CDS vs CDS map, has only one region of difference between the lineage at the reference at 262, 000 bp. This difference is clear in figure 3.33 where the nucleotide-level differences are given and then further in the zoomed map in 3.30. In the region is just one gene TETR0C001310 which is annotated as hypothetical, but it has 99.84% identity, 100% query cover to POL3 (YDL102W). This is a DNA polymerase subunit required for replication and double strand break repairs. There are also 12 predicted coding regions that are in areas of the Ethanol Red chromosome for which none of the strains in the lineage have a large degree of similarity. TETR0C05070 has 100% similarity and coverage to HLJ1 (YMR161W) which is a Hsp40p chaperone and analogous to DnaJ of *E. coli*. TETR0C05080 has 99.61%identity, 97% query cover to helicase YRF1-6 (YNL339C) (Ambroset *et al.*, 2011). TETR0C05090 is annotated as YBL111C a telomere-encoded helicase like protein (Shiratori *et al.*, 1999). TETR0C05100 is YBL112C of unknown function. TETR0C05110 is an uncharacterized retrotransposon like protein and finally TETR0C05120 another protein similar to gag-pol fashion proteins. These genes can however be found by BLASTN search elsewhere on the genome, as such these differences are likely to be assembly artefact and not contributing to any phenotype.

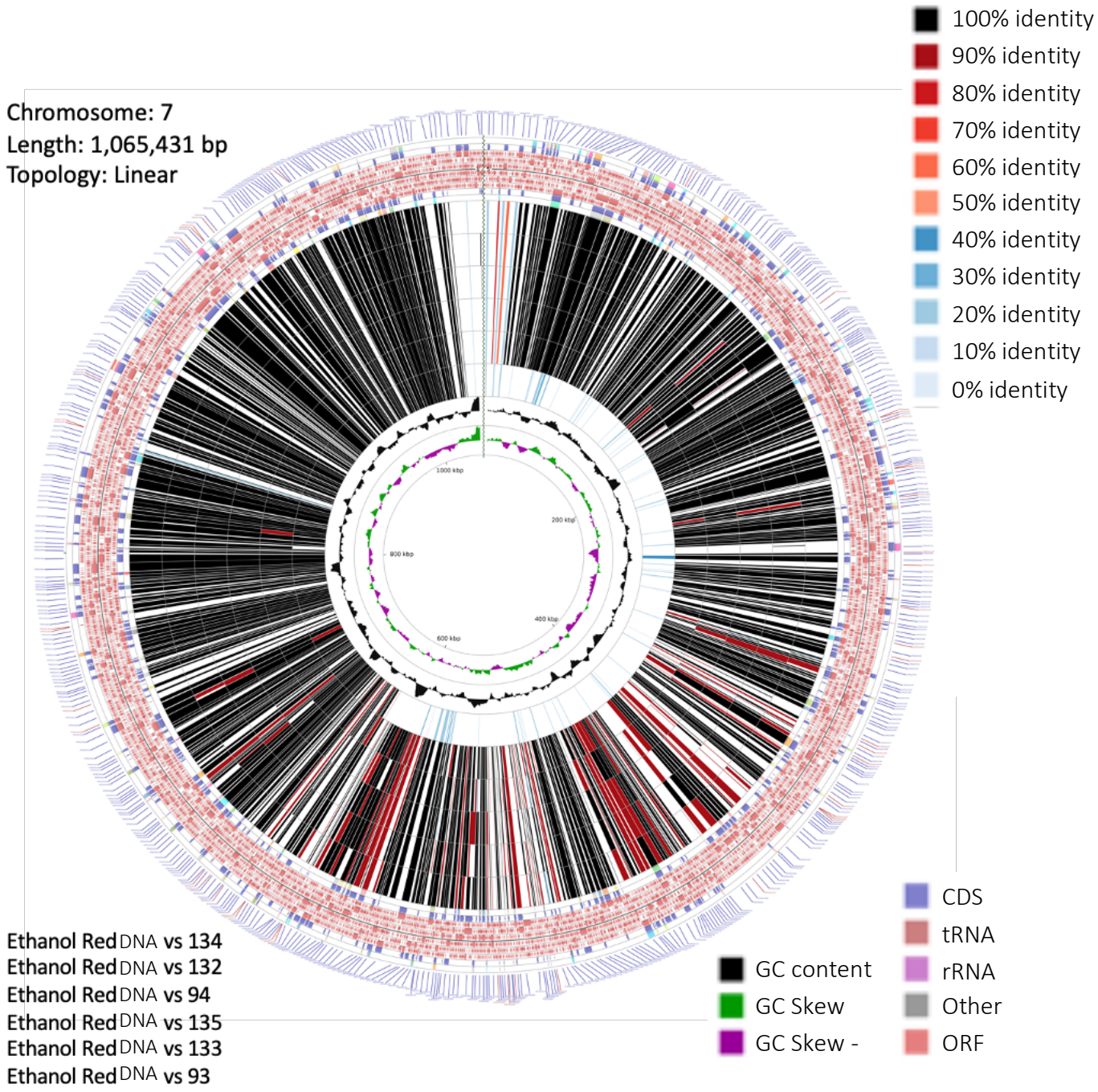


Figure 3.28 CCT map comparing the EtOH Red chromosome 7 with chromosome 7 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

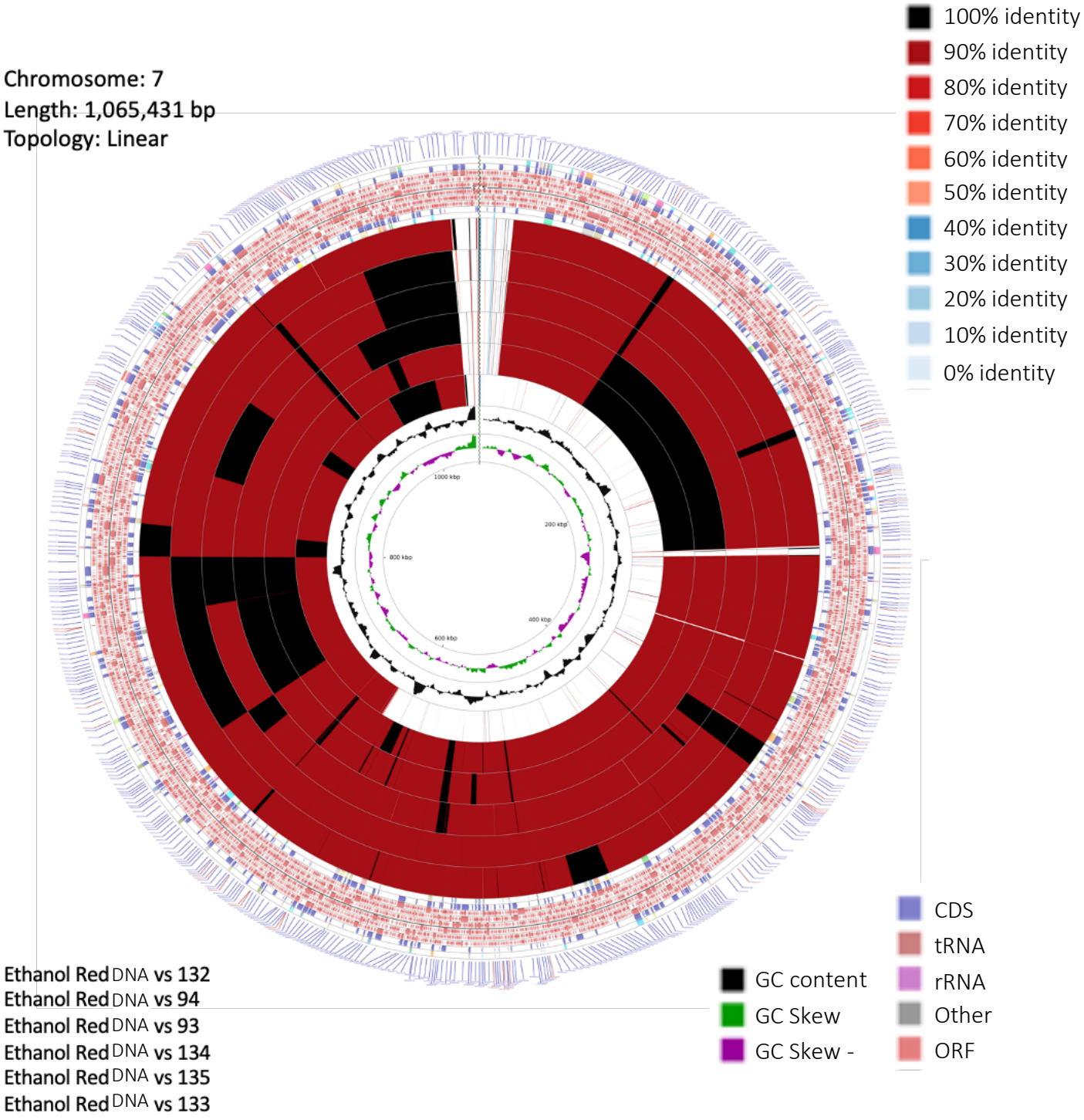


Figure 3.29 CCT map comparing the EtOH Red chromosome 4 with chromosome 4 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

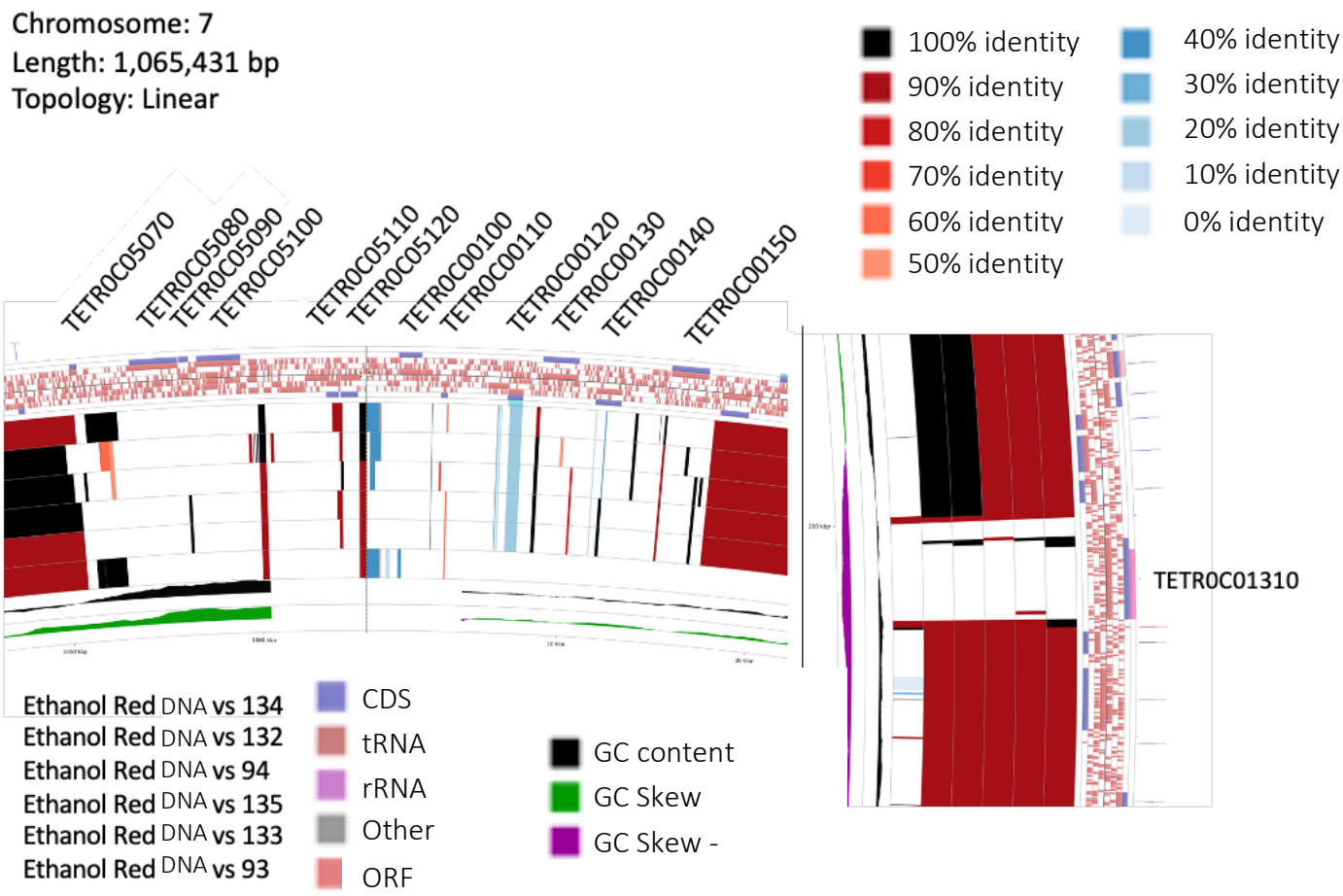


Figure 3.30 Zoomed CCT map of Chromosome 7. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.14 and 3.15. A) is the 0 bp region zoomed by a factor of 10. B) is the 262,000 bp region zoomed by a factor of 10.

Chromosome 8 once again has a very high degree of similarity across all of the strains in the lineage compared to Ethanol Red. This is evident in both the CDS vs CDS maps and the DNA vs DNA given in figures 3.31 and 3.32 respectively. The one region of difference is again at the telomeres which is shown in more detail in figure 3.33. This shows that there are 9 predicted coding regions in Ethanol Red that are missing from all strains in the lineage. TER0L02640 has 100% identity and query coverage to PAU10 (YDR542W) (Luo and van Vuuren, 2009). TER0L02650 is annotated as YAR068W which is of unknown function but has been found to be induced in certain stress conditions (Nishida, 2006). TER0L02670 is annotated as a protein of unknown function but has 100% identity over 94% of the query sequence to Pho12p which is an acid phosphatase (Venter and Hörz, 1989). TER0L02680 is IMD4, inosine monophosphate dehydrogenase which is involved in nucleotide synthesis at a rate limiting step and as such is important for cellular growth and would impact growth rate (Escobar-Henriques and Daignan-Fornier, 2001). TER0L01100 has 86.57% identity over 100% of the protein sequence query to Mal63 which is a regulator involved in maltose utilisation (Higgins *et al.*, 1999). TER0L02660, TER0L02700 and TER0L00100 have no similarity to any protein in the databases search. In figure 3.33 it can be observed that whilst there is no identity to the coding sequence of these genes in the lineage there are areas of high level of identity more proximal to the telomeres than the coding sequence. This suggests that in this case, the later of the two explanations offered earlier for the lack of identity to telomeric encoded proteins in the lineage may be true in this instance. Reduced sequencing coverage at the telomeres doesn't adequately explain this pattern of identity making a recombination event a more sufficient explanation. Of these genes, most are found in the genome elsewhere using BLAST (other than Pho12p). This could still be an artefact of the assembly, only targeted sequencing by Sanger or similar would be definitive here.

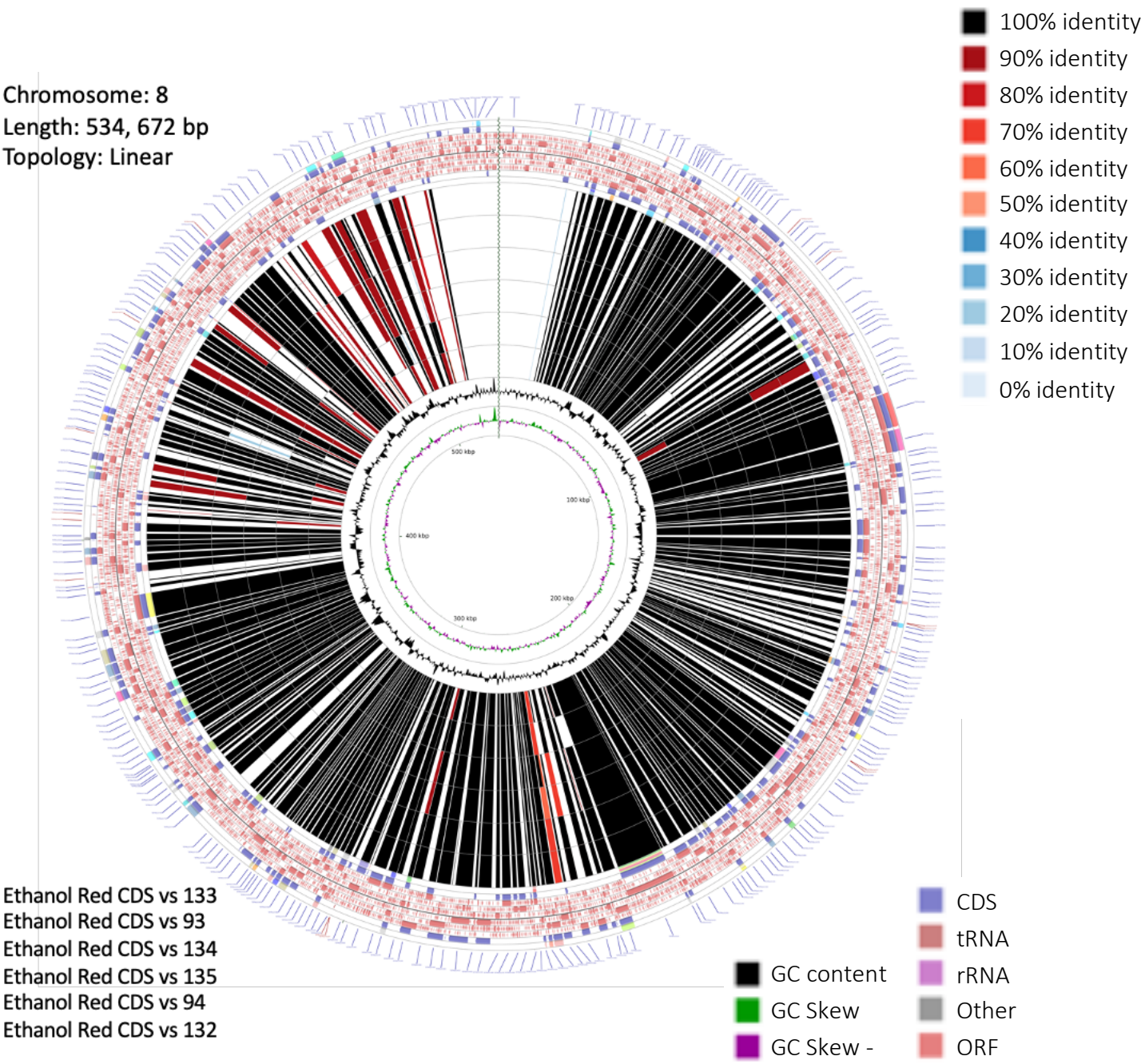


Figure 3.31 CCT map comparing the EtOH Red chromosome 8 with chromosome 8 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

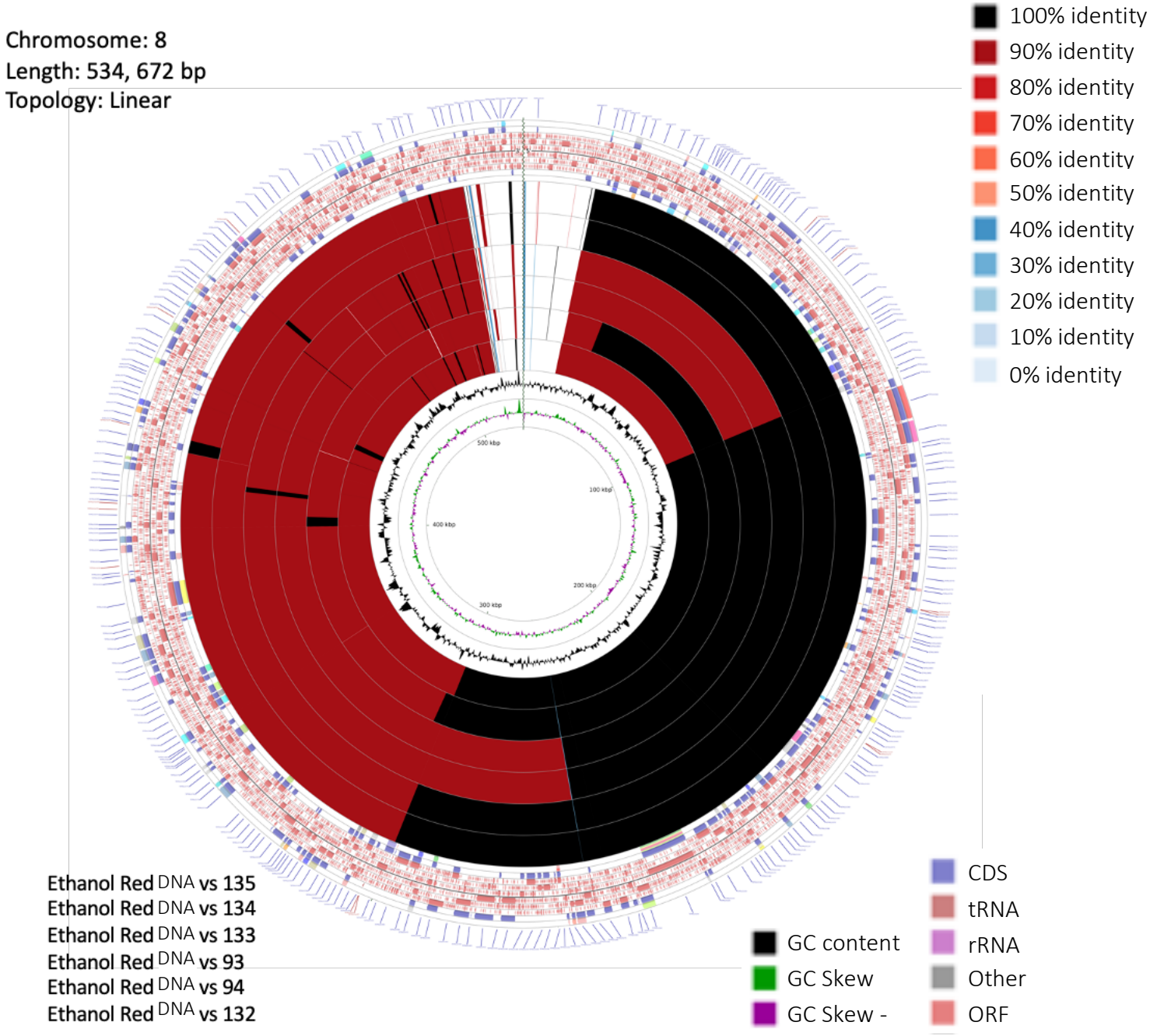


Figure 3.32 CCT map comparing the EtOH Red chromosome 8 with chromosome 8 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

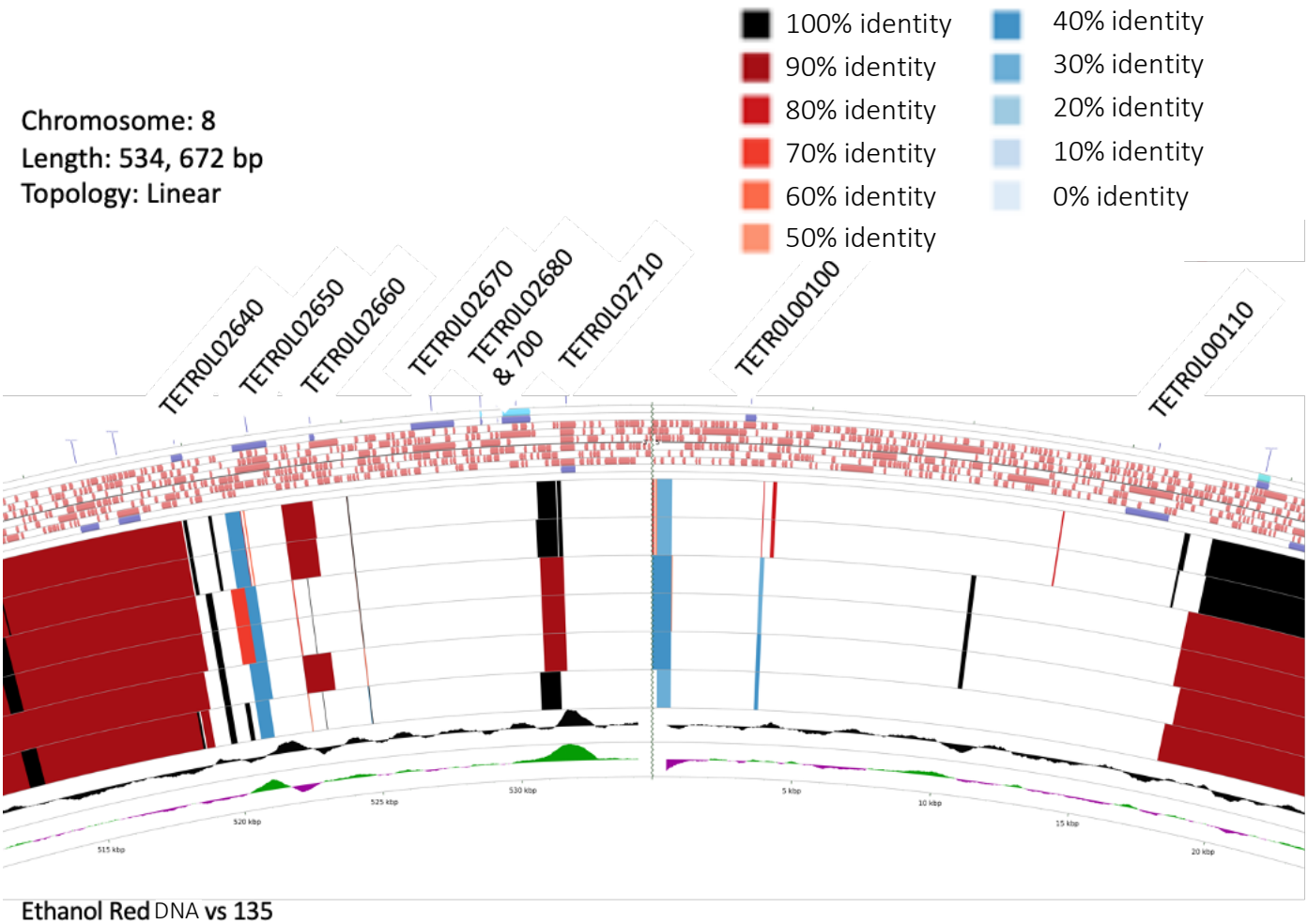


Figure 3.33 Zoomed CCT map of Chromosome 8. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.14 and 3.15. Shown is the 0 bp region zoomed by a factor of 6.

Chromosome 9 of 437, 035 bp (figure 3.34) only differs at the end of chromosome and similar to others all of the strains have the same differences as each other when compared to the reference. This difference is clearly visible in the DNA vs DNA map in figure 3.35 and highlighted in figure 3.36. In this case the region of difference contains 10 predicted coding regions. TETR0M02120 has 100% similarity by BLASTN search to hexose transporter HXT11 (YOL156W) (Boles and Hollenberg, 1997). TETR0M02130 has 100 % similarity to IMA3 isomaltase.TETR0M02140 is likely a miss-annotation as it predicts a protein of only 5 amino acids long. TETR0M02150 which shows 99.4% identity over 99% query cover to putative membrane glycoproteins, VTH2 (Cooper and Stevens, 1996). TETR0M02160 is YNR075W which is involved in plasma membrane protein organisation (Spode *et al.*, 2002). TETR0M02170 is again PAU6 (YNR076W) similar to regions of difference in other chromosomes. TETR0M02180 and TETR0M02190are YEL076C-A and YEL076C respectively which are both of unknown function. TETR0M02100 which shows 97.69%identity, 100% query cover to YRF1-8 (YOR396W) is a telomere encoded helicase (Yamada *et al.*, 1998). Finally TETR0M02110 was annotated as a hypothetical protein but has similarity to several entries in the NCBI database including HLJ1 (YMR161W) a co-chaperone for Hsp40p (Beilharz *et al.*, 2003). These genes are found in multiple other locations in the chromosome, potentially as a result of misassembly.

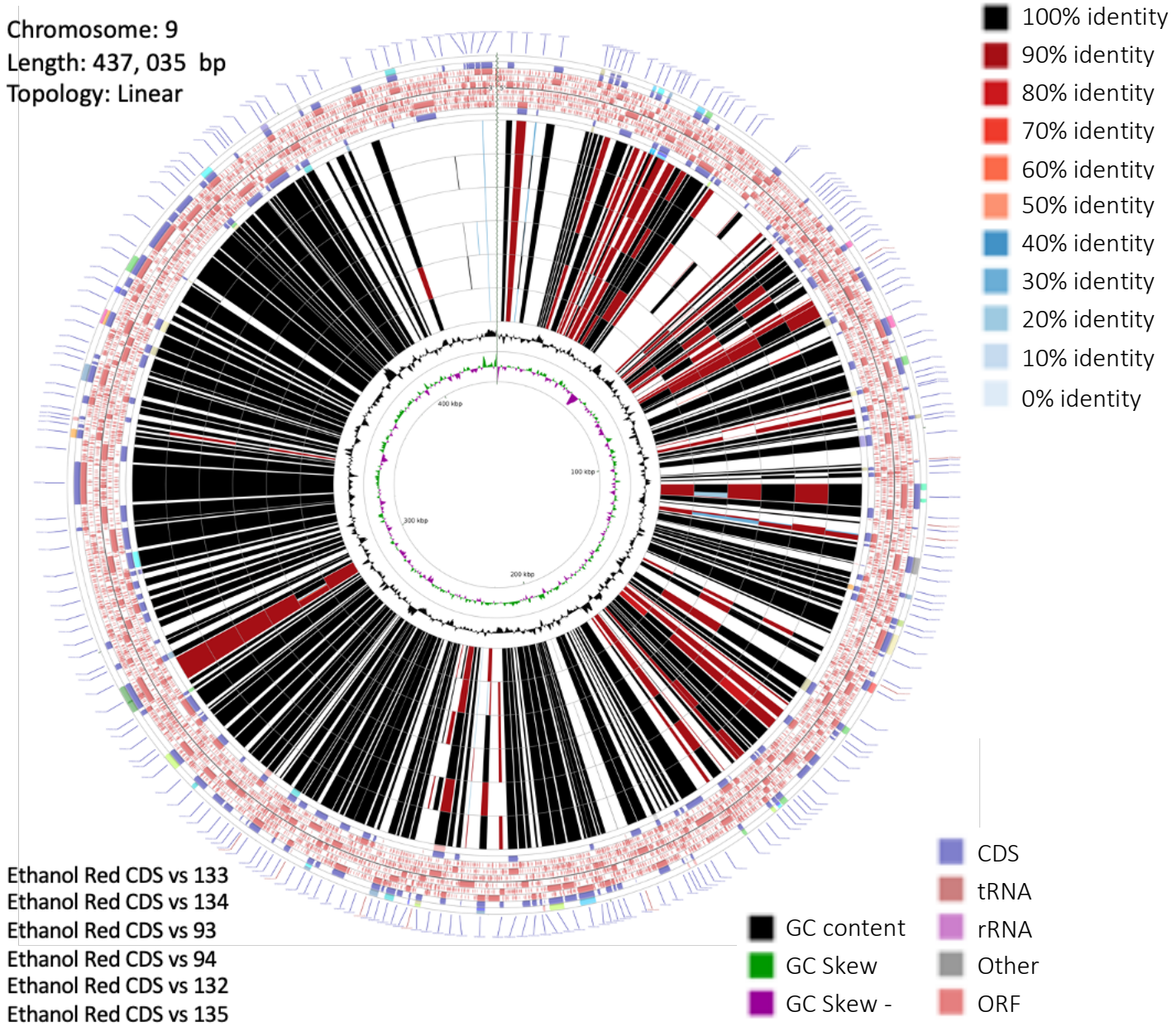


Figure 3.34 CCT map comparing the EtOH Red chromosome 9 with chromosome 9 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

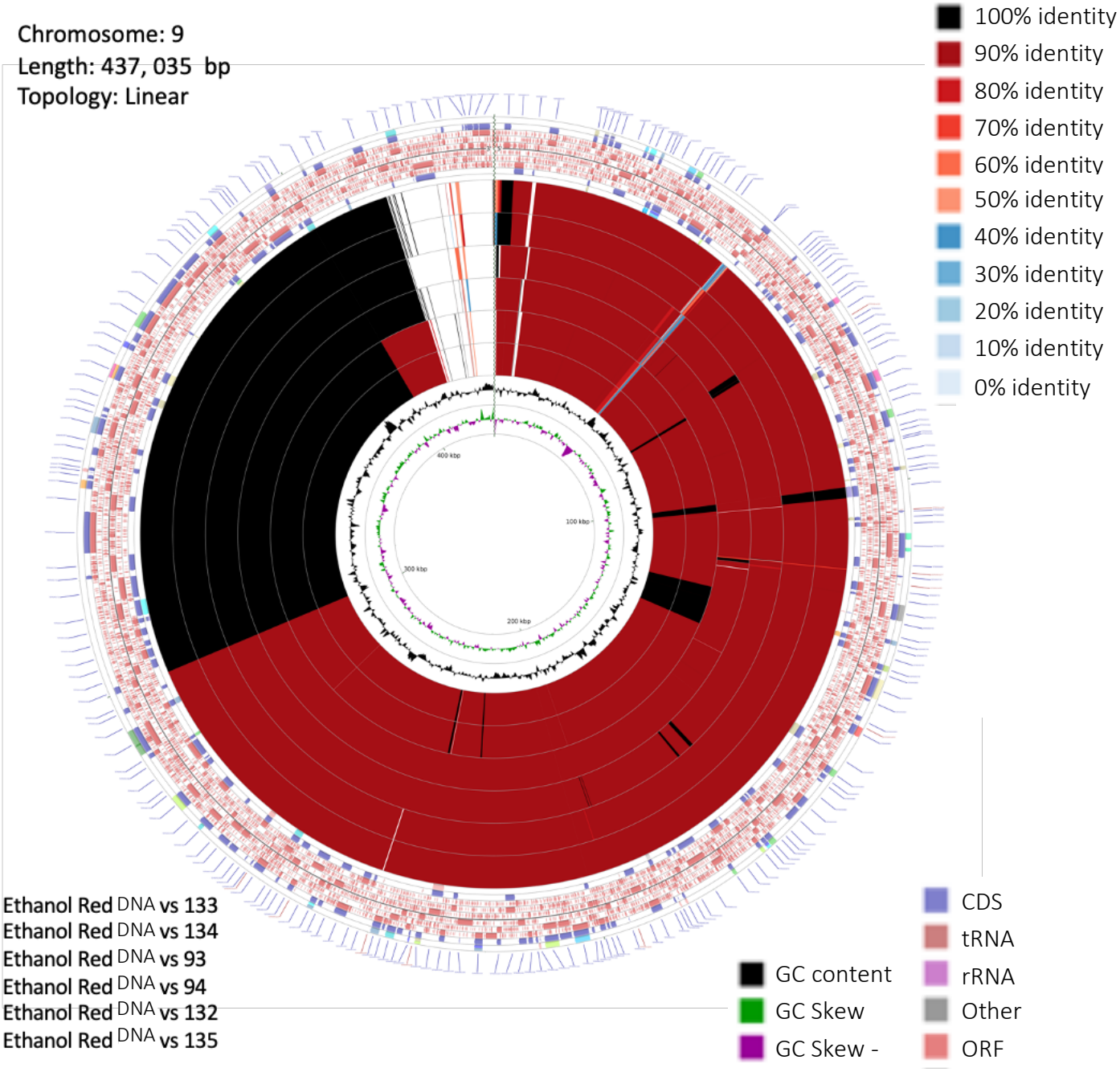


Figure 3.35 CCT map comparing the EtOH Red chromosome 9 with chromosome 9 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

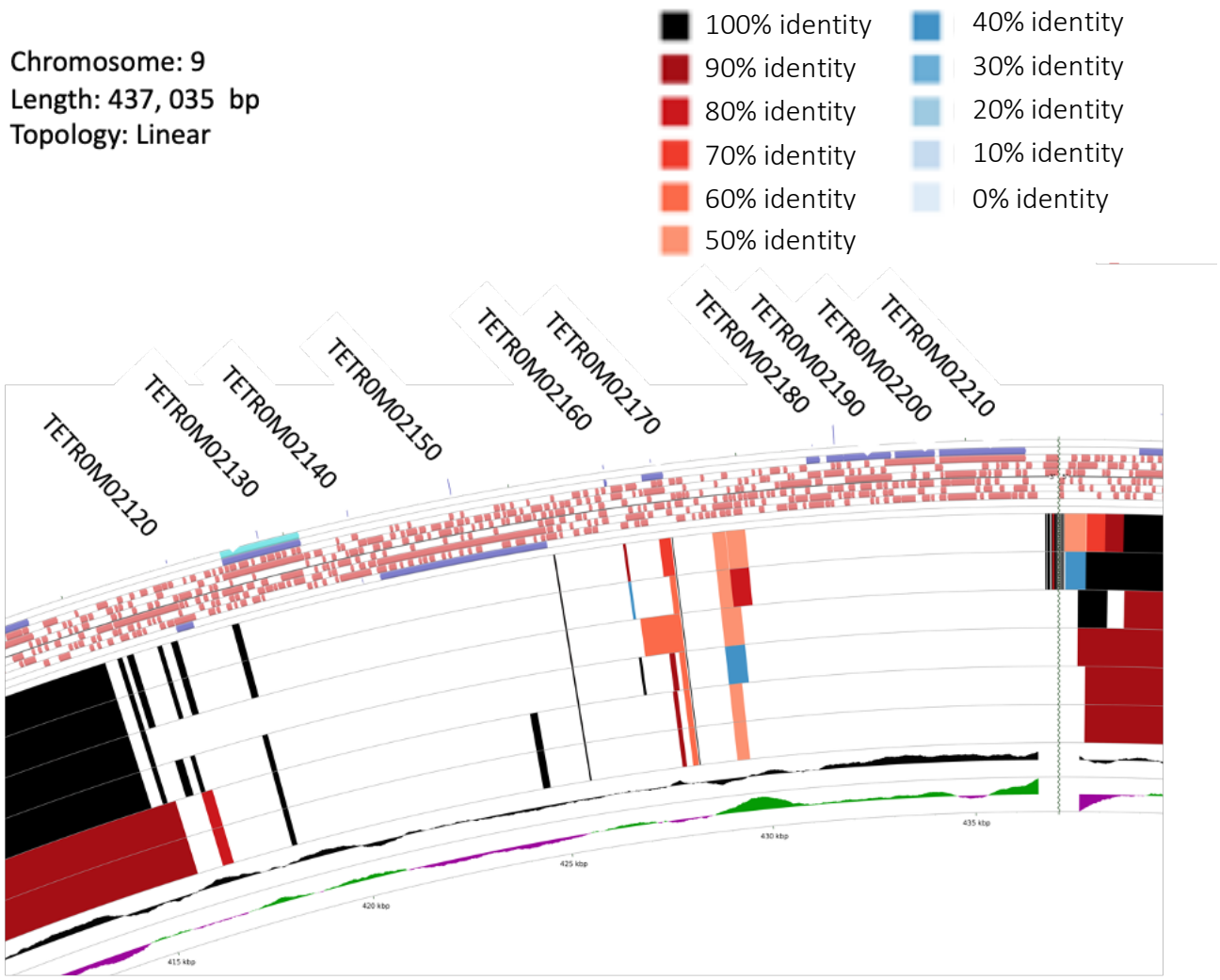


Figure 3.36 Zoomed CCT map of Chromosome 7. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.14 and 3.15. A) is the 425,000bp region zoomed by a factor of7.

Chromosome 10 has no large regions where there is a <80% similarity on either a CDS or DNA level between the lineage and the Ethanol Red reference. Again, as in chromosome one, there appears to be a considerable number of regions where CDS are present in the reference genome, but not comparator sequences, in the CDS map, but the DNA sequence appears to be conserved. This can be attributed to areas where there is no protein coding region predicted on the reference genome. Based on manual inspection of the short Illumina reads to their respective assemblies and the results from Ideel indicating truncated proteins, it is most likely that these inconsistencies come from errors in protein coding region prediction as well as some assembly error. This is suggested at the CDS level shown in figure 3.37 and further emphasised in the nucleotide level comparison in figure 3.38.

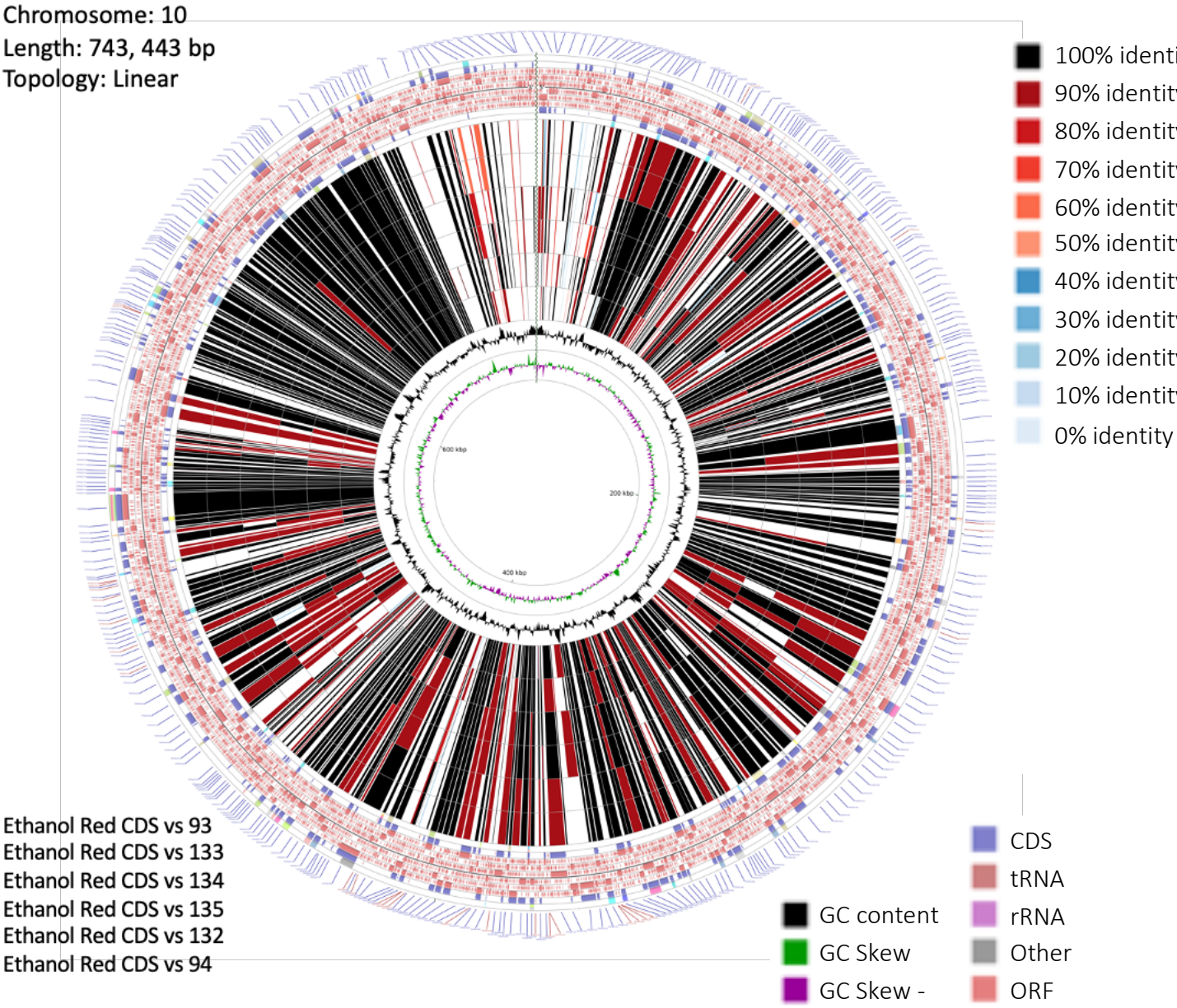


Figure 3.37 CCT map comparing the EtOH Red chromosome 10 with chromosome 10 from 6 other strains. This map was generated using CCT maps build_blast_atlas.shscript. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

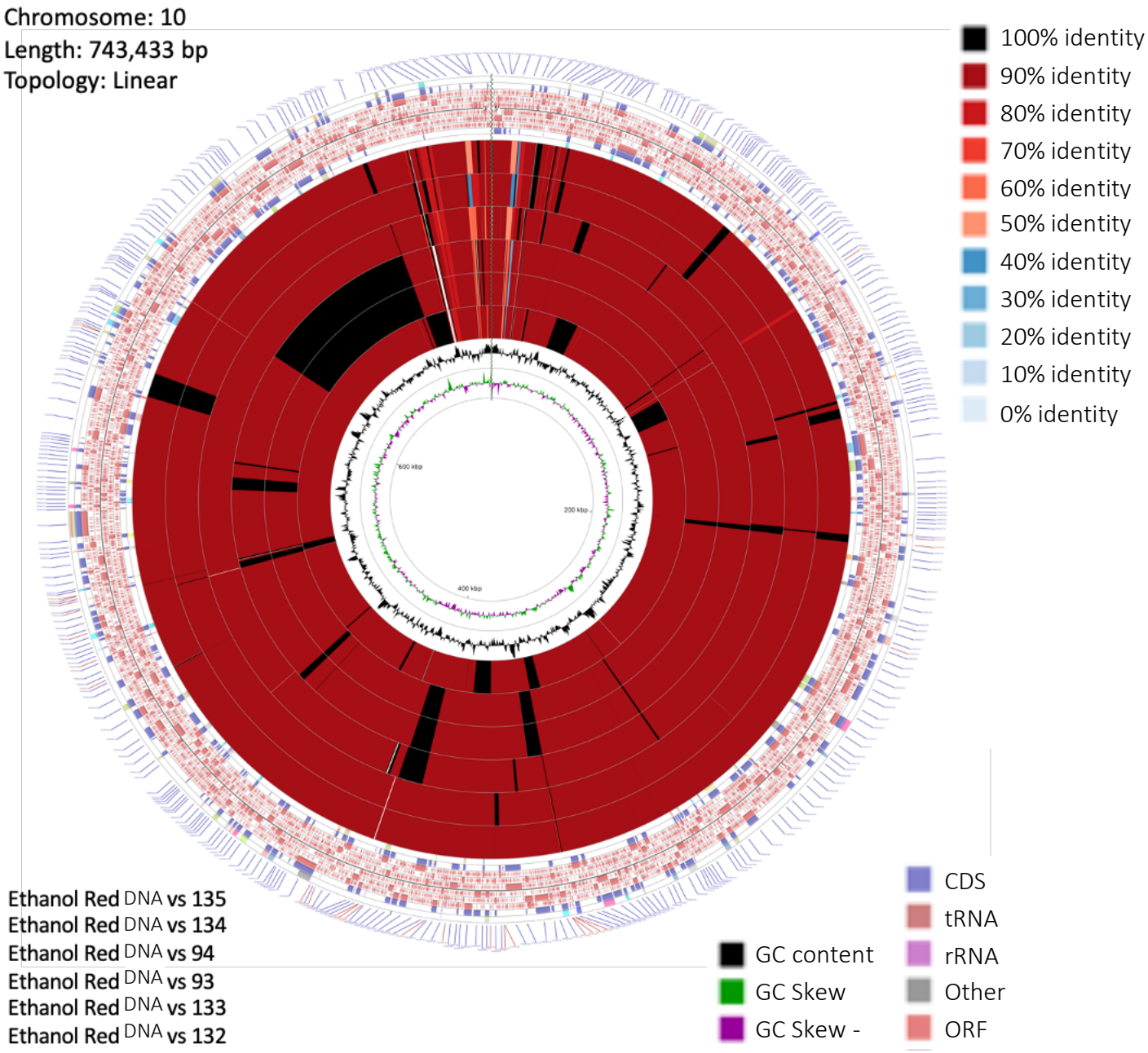


Figure 3.38 CCT map comparing the EtOH Red chromosome 4 with chromosome 4 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

Similarly, chromosome 11 has large degrees of similarity over the full chromosome in both CDS and DNA as can be seen in figures 3.39 and 3.40 respectively. Though there does appear to be a small truncation or reduced sequencing coverage at one telomere in the lineage compared to Ethanol Red in a non-coding sequence. This type of situation which is seen in a number of chromosomes is suggestive that the availability of long read nanopore sequence reads for Ethanol Red has allowed for improved telomere sequencing coverage of terminal regions in particular compared to short-read only assemblies even when the short-read assemblies have a complete genome as reference.

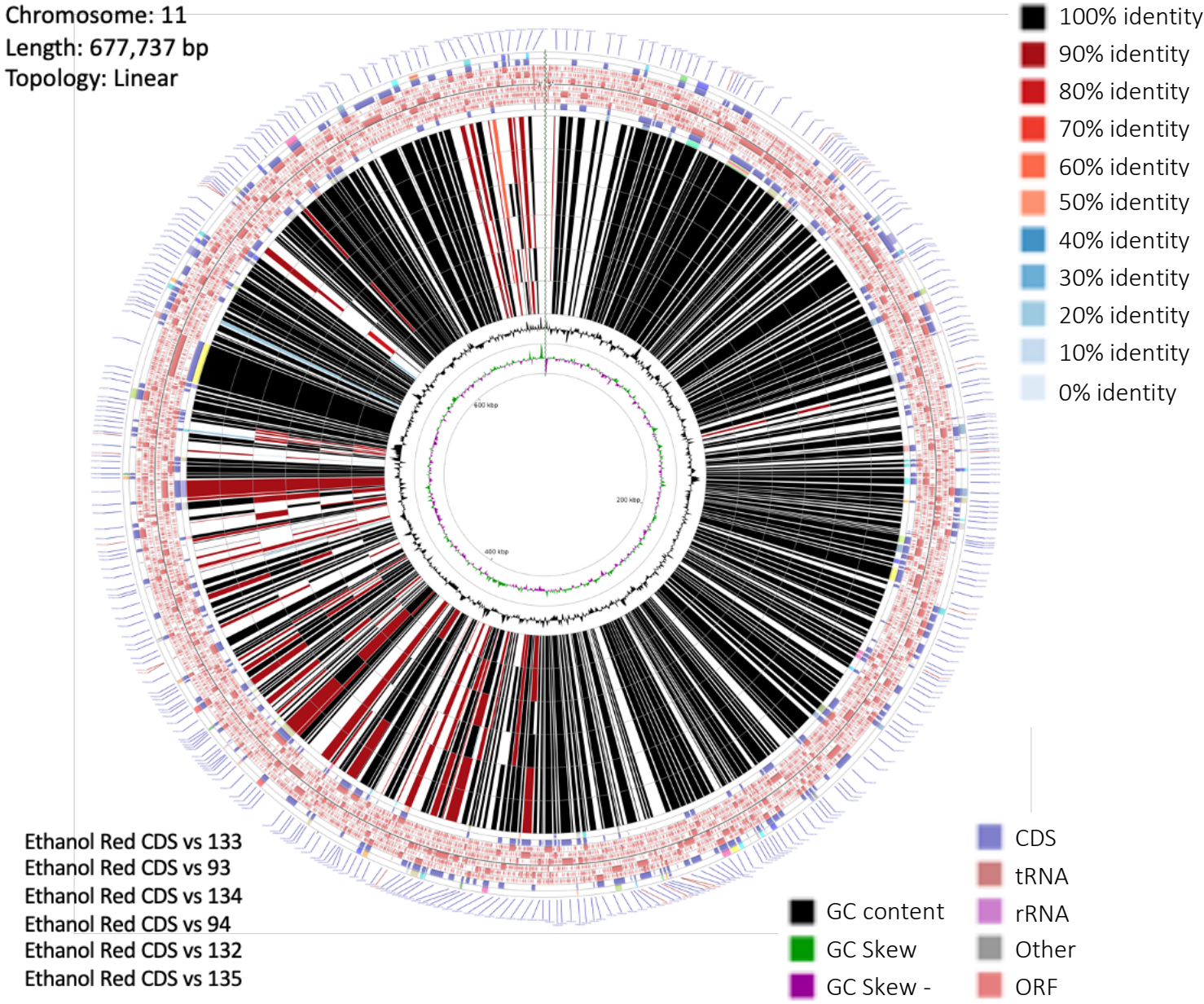


Figure 3.39 CCT map comparing the EtOH Red chromosome 11 with chromosome 1 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh_script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

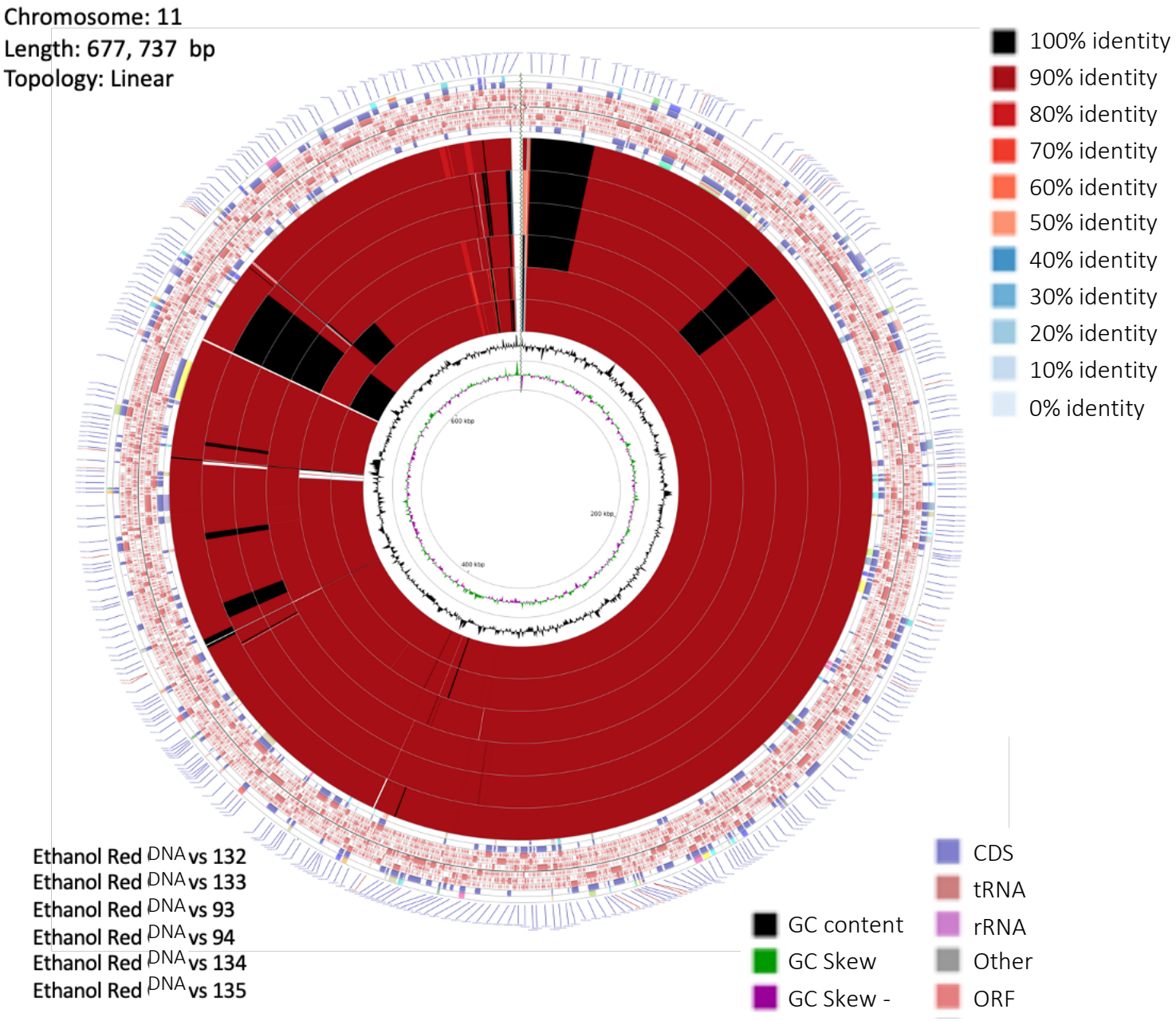


Figure 3.40 CCT map comparing the EtOH Red chromosome 4 with chromosome 4 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

Chromosome 12 has two regions where there appears to be large deletions in one of multiple lineage strains compared to the reference. Outside of these areas, the identity remains consistently high. Genome 133 which has a no/little similarity over the last 200 kbp of the reference genome, when checked for sequence length of the scaffold 133 was ~200kbp shorter than Ethanol Red. Suggesting that there has been an assembly issue, resulting in the loss of this section of the sequence. Similarly, the region around 600 kbp that appears to be largely absent instrains 93, 134 and 135 is approximately 100 kbp which corresponds to the difference in sequence length in these cases too. This suggests that the assembly of this chromosome has not been fully successful in 4 out of the 6 lineage strains rather than the chromosomes being genuinely shorter.

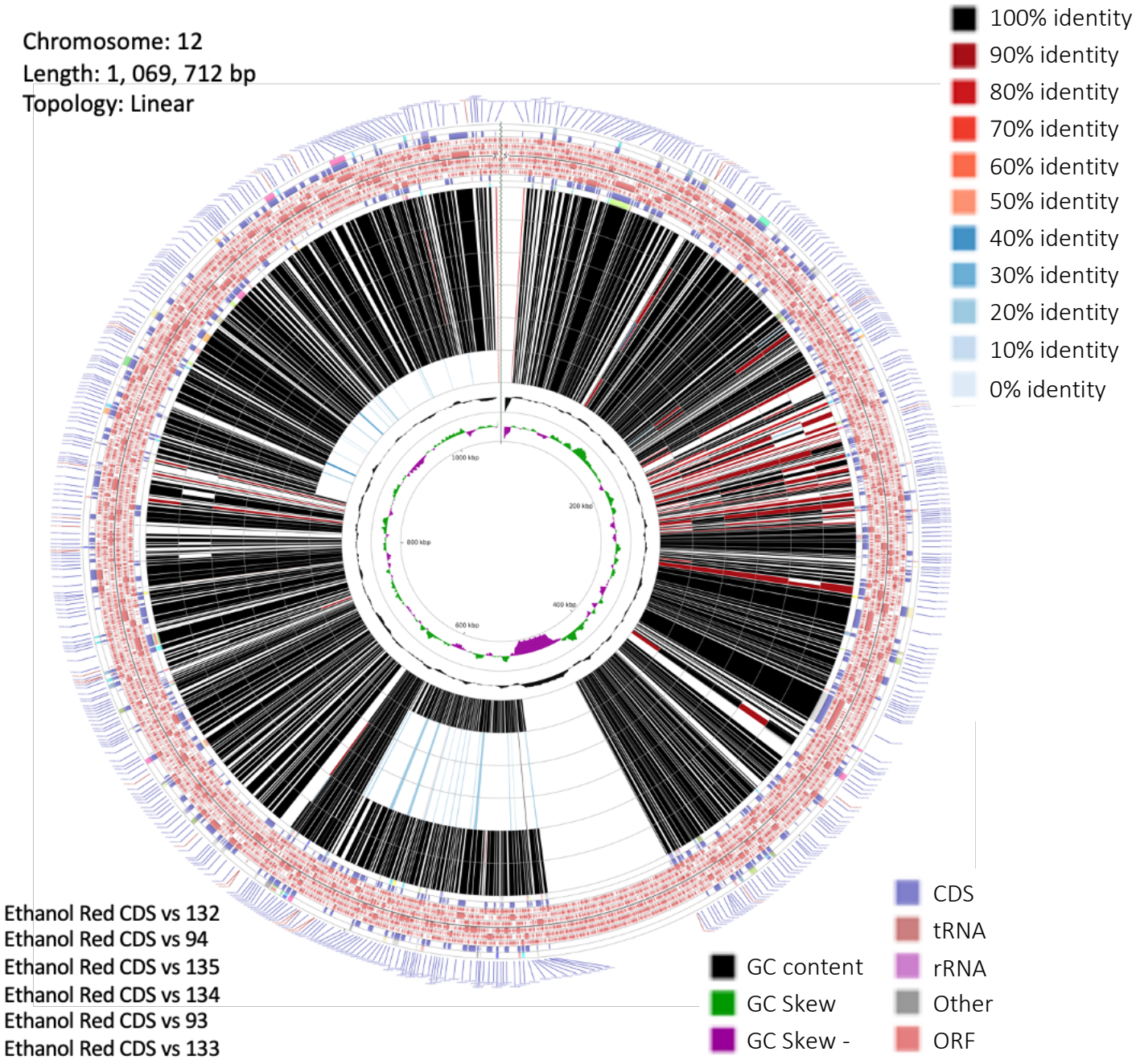


Figure 3.41 CCT map comparing the EtOH Red chromosome 12 with chromosome 12 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh_script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

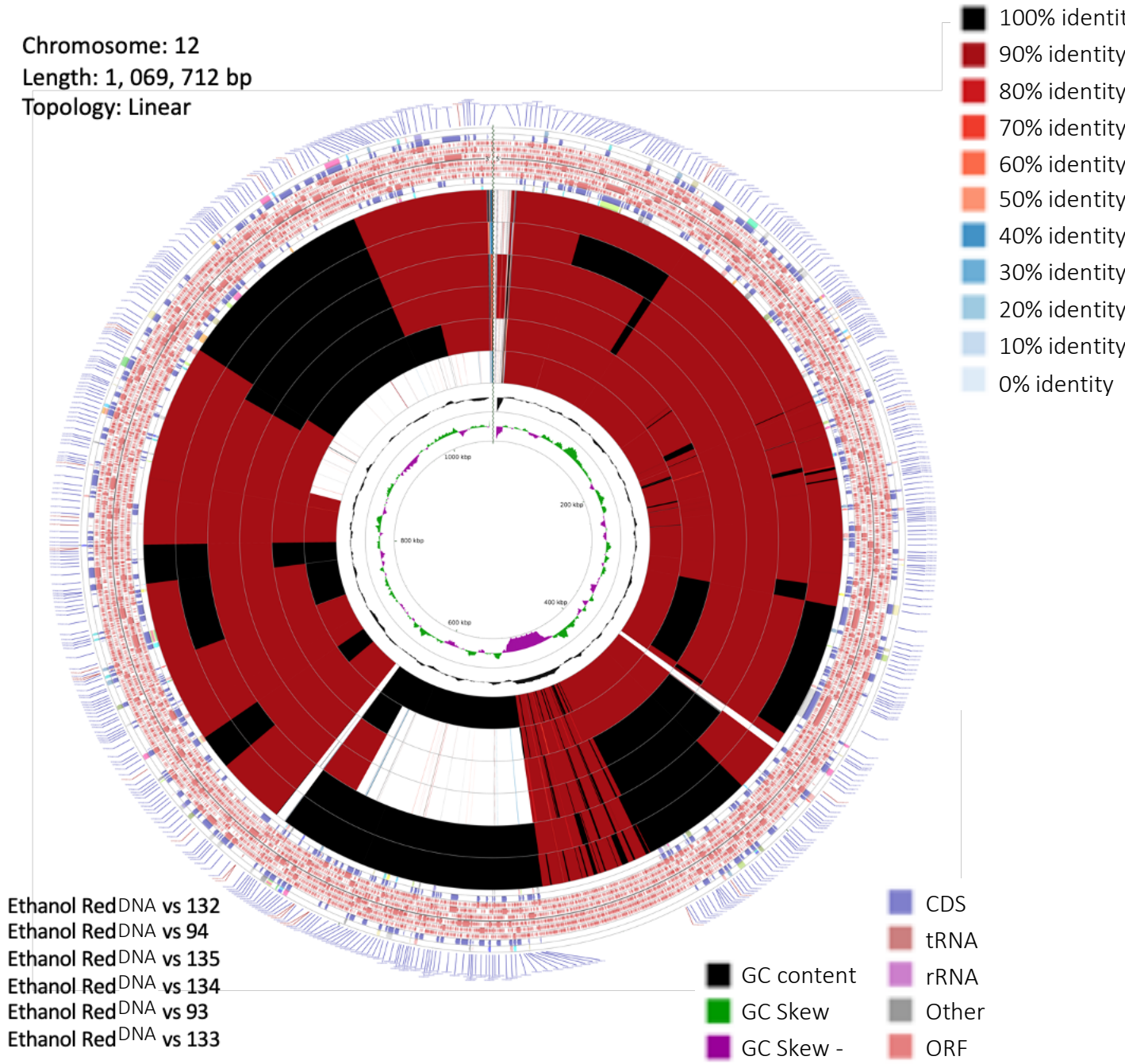


Figure 3.42 CCT map comparing the EtOH Red chromosome 12 with chromosome 12 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

Chromosome 13 only has one predicted coding region which is present in Ethanol Red but not all of the strains of the lineage. In this case the gene is only present in 94, 134, 135, 132 but not 133 and 93, due to its proximity to the telomeric region this is likely due to reduced sequencing coverage and subsequent missasembly In figure 3.43 which shows the CDS vs CDS and figure 3.44 which shows the DNA vs DNA the general similarity can be seen. The zoomed map in figure 3.45 shows the region of difference in more detail and contains three predicted coding regions. TETR0E04440 is a protein of unknown function annotated as YMR317W. TETR0E04450 is ADH6 (YMR318C) ADH6, an alcohol dehydrogenase (Larroy *et al.*, 2002) and finally TETR0E04460 FET4 (YMR319C) which is responsible for Iron transport (Dix *et al.*, 1994). There is also another, small, predicted coding region in Ethanol Red chromosome five that is not present in all the lineage strains. This is also shown in figure 3.45 and is not present in strain 132, it is annotated as TETR0E04470 which is YMR321C, a protein of unknown function but with similarity to some membrane proteins.

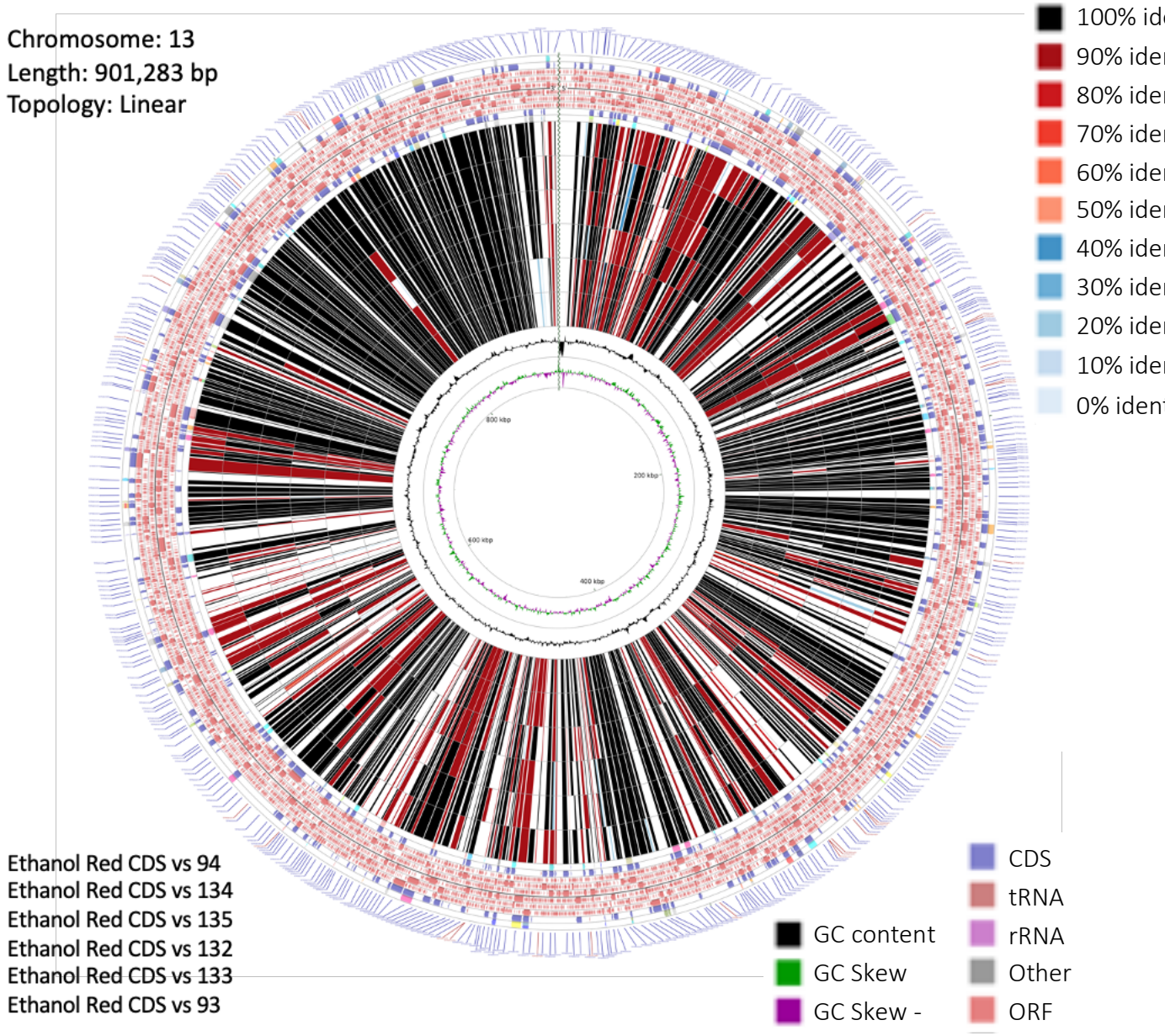


Figure 3.43 CCT map comparing the EtOH Red chromosome 13 with chromosome 13 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

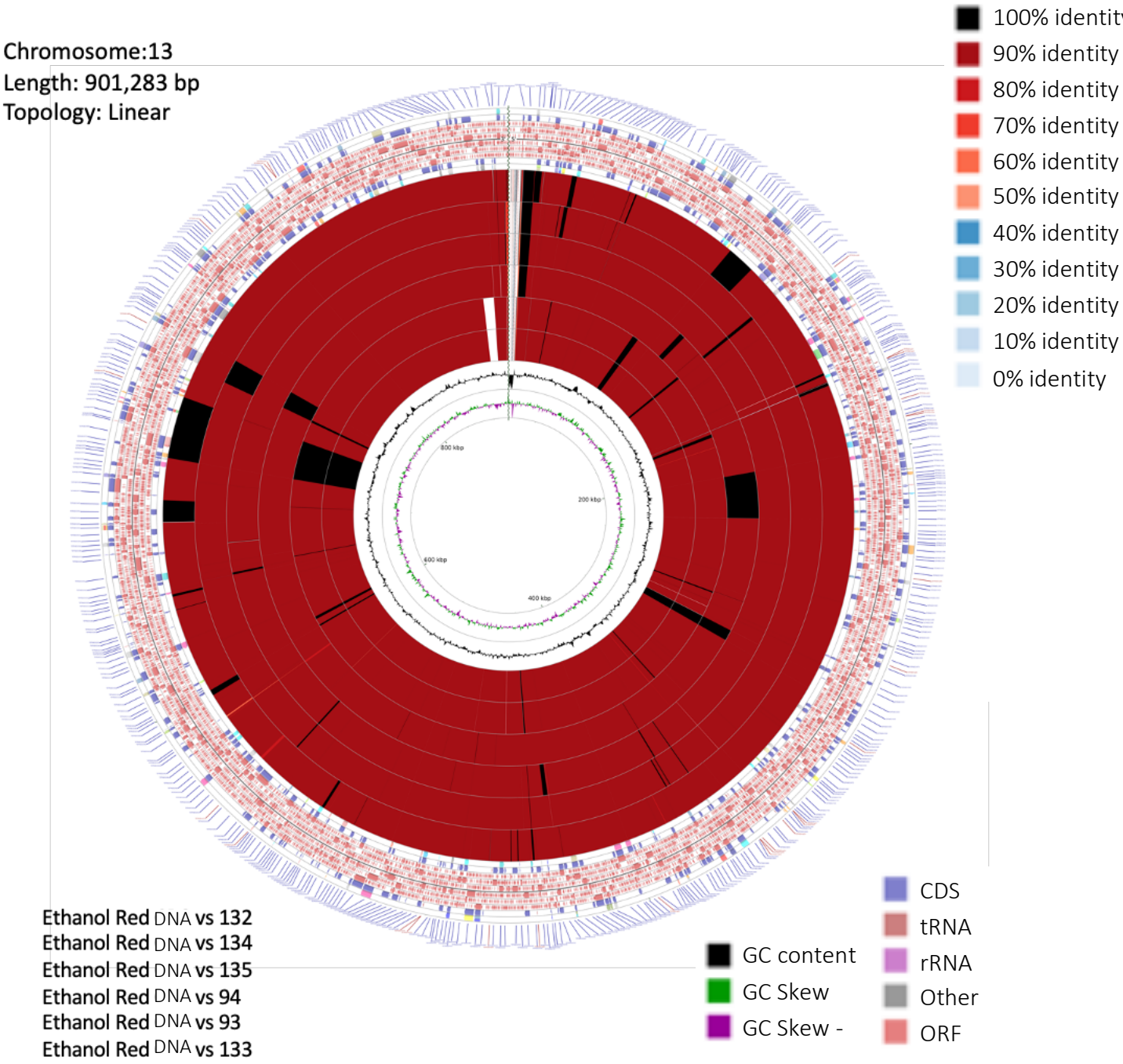


Figure 3.44 CCT map comparing the EtOH Red chromosome 13 with chromosome 13 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

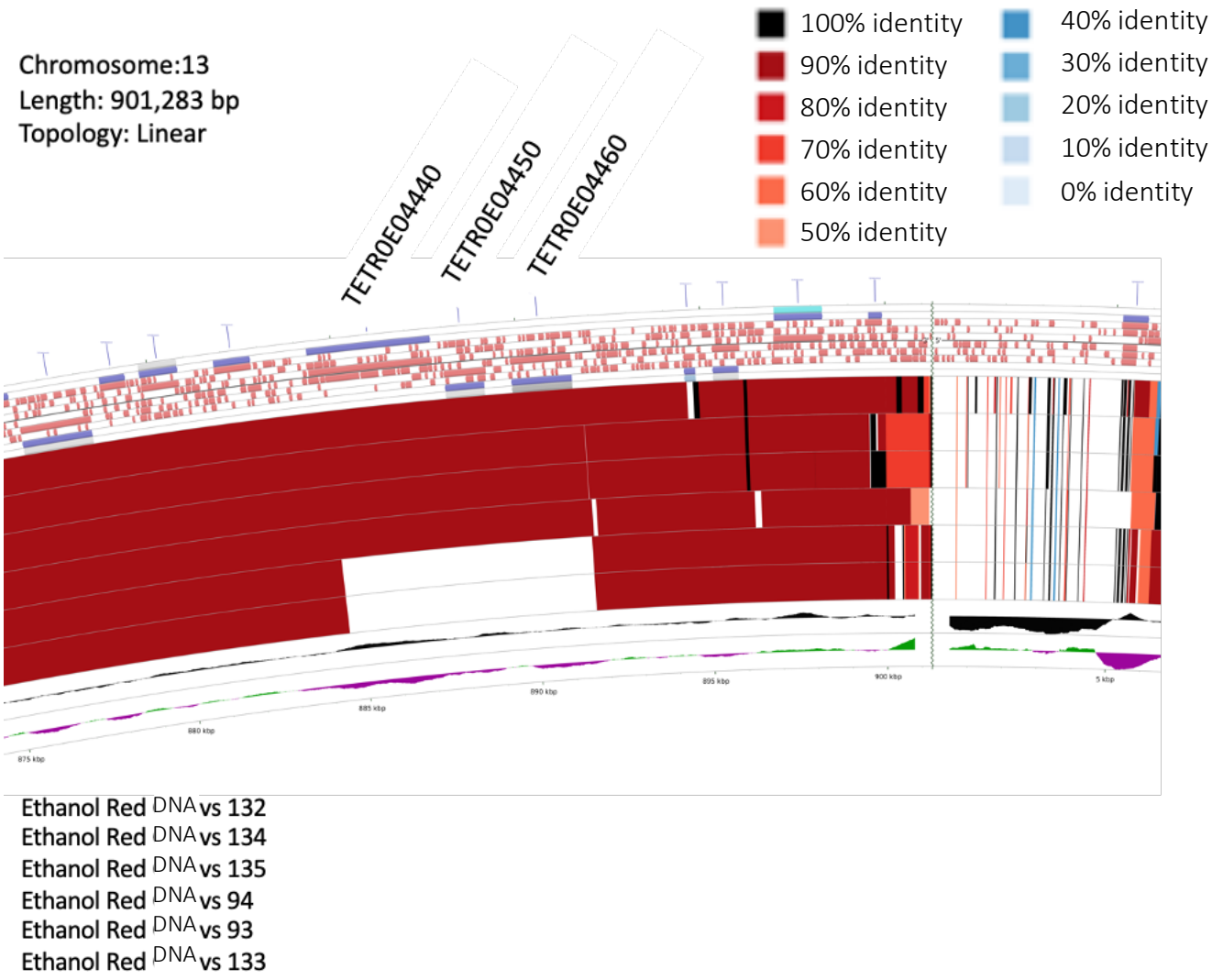


Figure 3.45 Zoomed CCT map of Chromosome 13. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.43 and 3.43. Shown is the 89-,000 bp region zoomed by a factor of 12.

Chromosome 14 shown in figure 3.46 has almost full identity over the protein-coding region and this is also seen at a nucleotide level in figure 3.47. The two regions where all strains are missing protein-coding sequences from the reference are shown in figure 3. 48. TETR0H03620 which has 100% identity and query cover to a gag-pol fusion protein. TETR0H03630 and TETR0H03640 are also most similar to gag-pol fusion proteins but to a slightly lesser extent with 98% and 99% identity respectively. TETR0H00230 and TETR0H00240 have no similarity to any protein of known function. There seems to be reduced instance of gag-pol proteins in the lineage strains compared to the reference. Given that these are required for complete viral packaging into a capsid, this is suggestive that the lineage strains be more resistant to effects of viruses than the reference (Fujimura *et al.*, 1992). There is also a small region where strain 133 and 134 are missing six consecutive predicted protein-coding regions. Three adjacent proteins coding sequences TETR0H01960 (MIC27), TETR0H01970 (AVT4) and TETR0H01980 (POL1) are all mitochondrial related genes with involvement in membrane structure and interactions and DNA synthesis (BURGERS *et al.*, 1990; Russnak, Konczal and McIntire, 2001; von der Malsburg *et al.*, 2011). Finally, on this telomere is TETR0H01990 is MET4 which is a regulator of amino acid pathways. On the other telomere are TETR0H01100 which is LEU4 for leucine biosynthesis (Baichwal *et al.*, 1983) and TETR0H01110 (INP52) which is involved in various cell signalling pathways (Stolz *et al.*, 1998). These genes are found in multiple other locations in the chromosome, potentially as a result of misassembly.

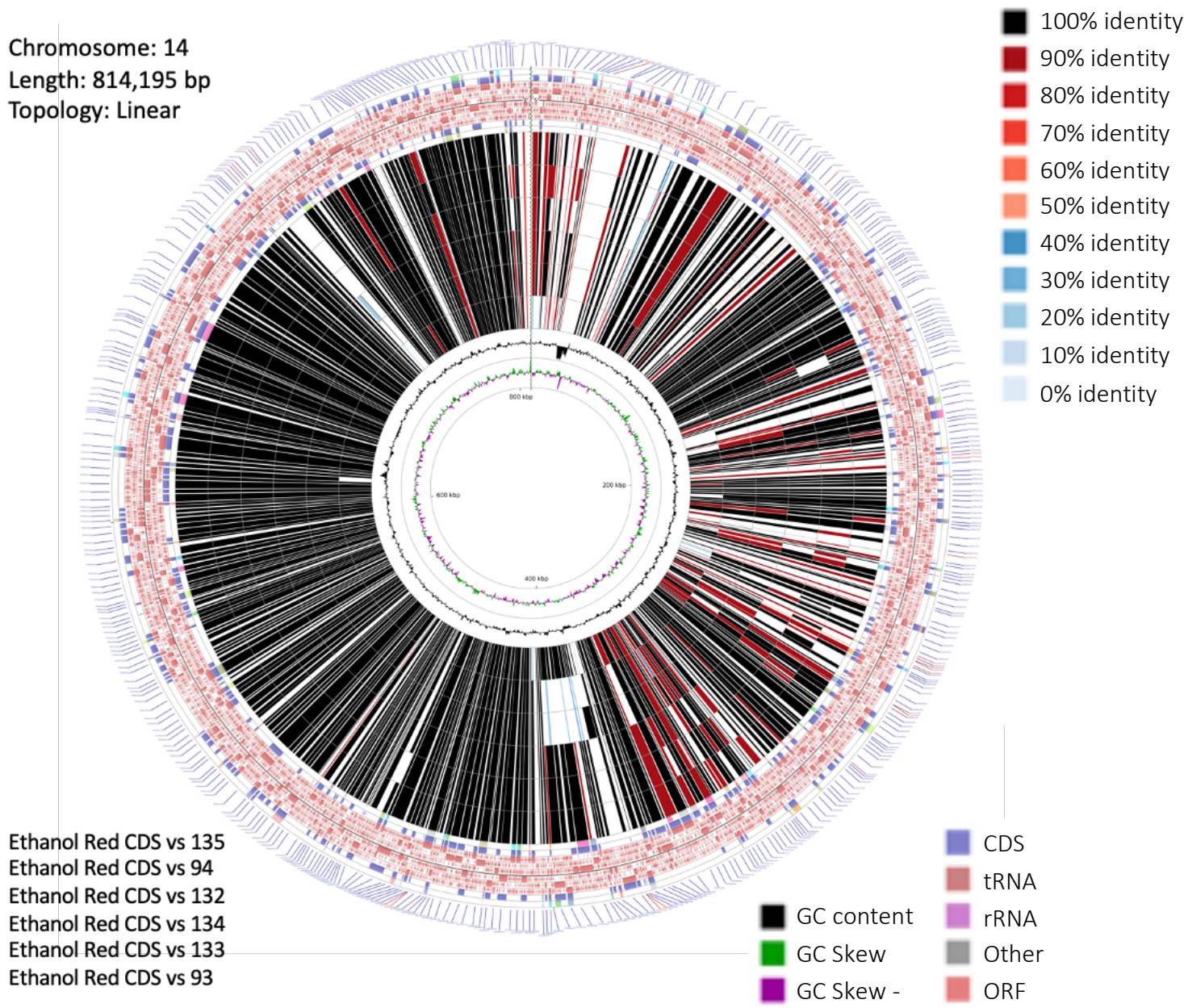


Figure 3.46 CCT map comparing the EtOH Red chromosome 7 with chromosome 7 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

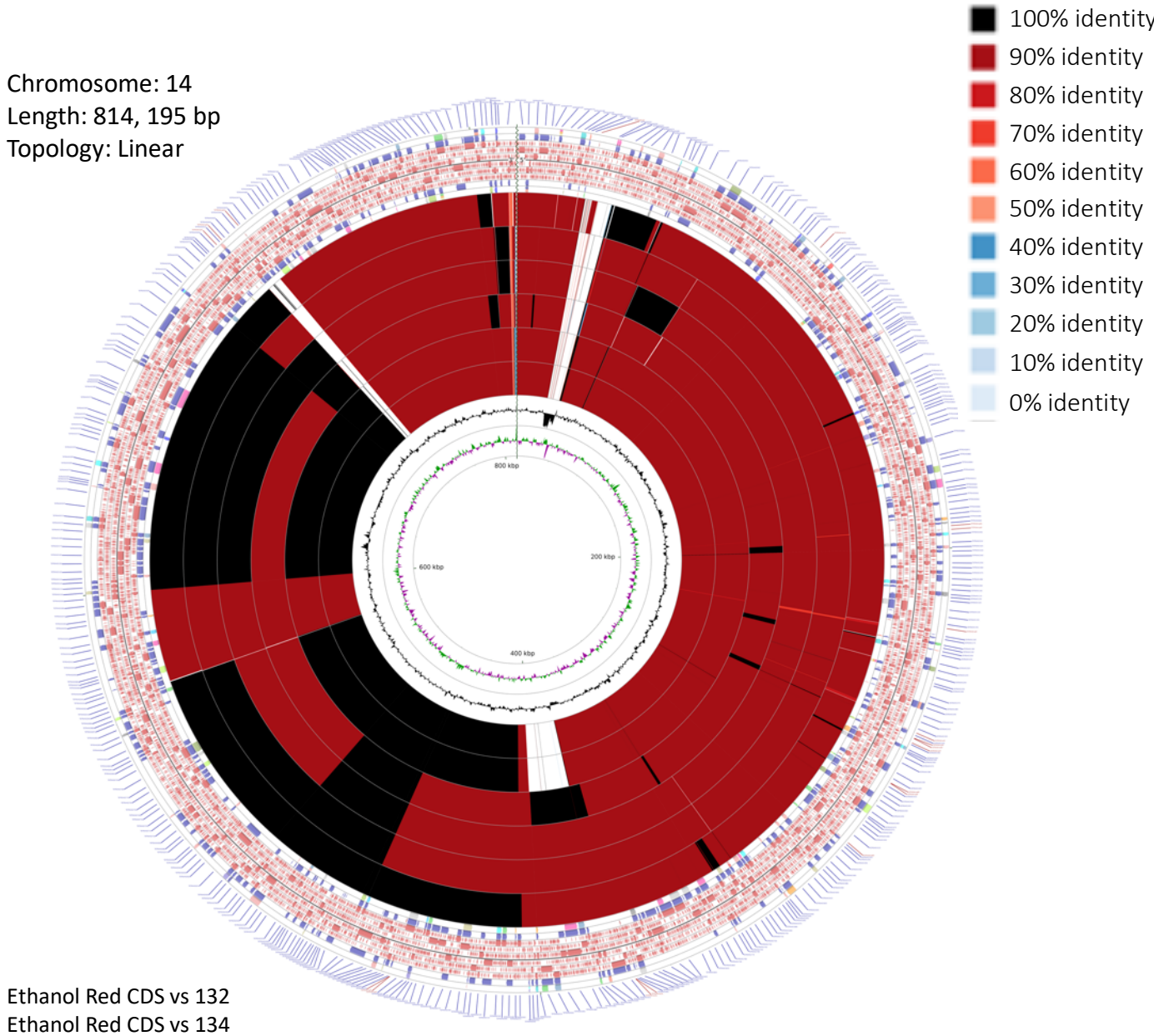


Figure 3.47 CCT map comparing the EtOH Red chromosome 14 with chromosome 14 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

Chromosome: 14
Length: 814, 195 bp
Topology: Linear

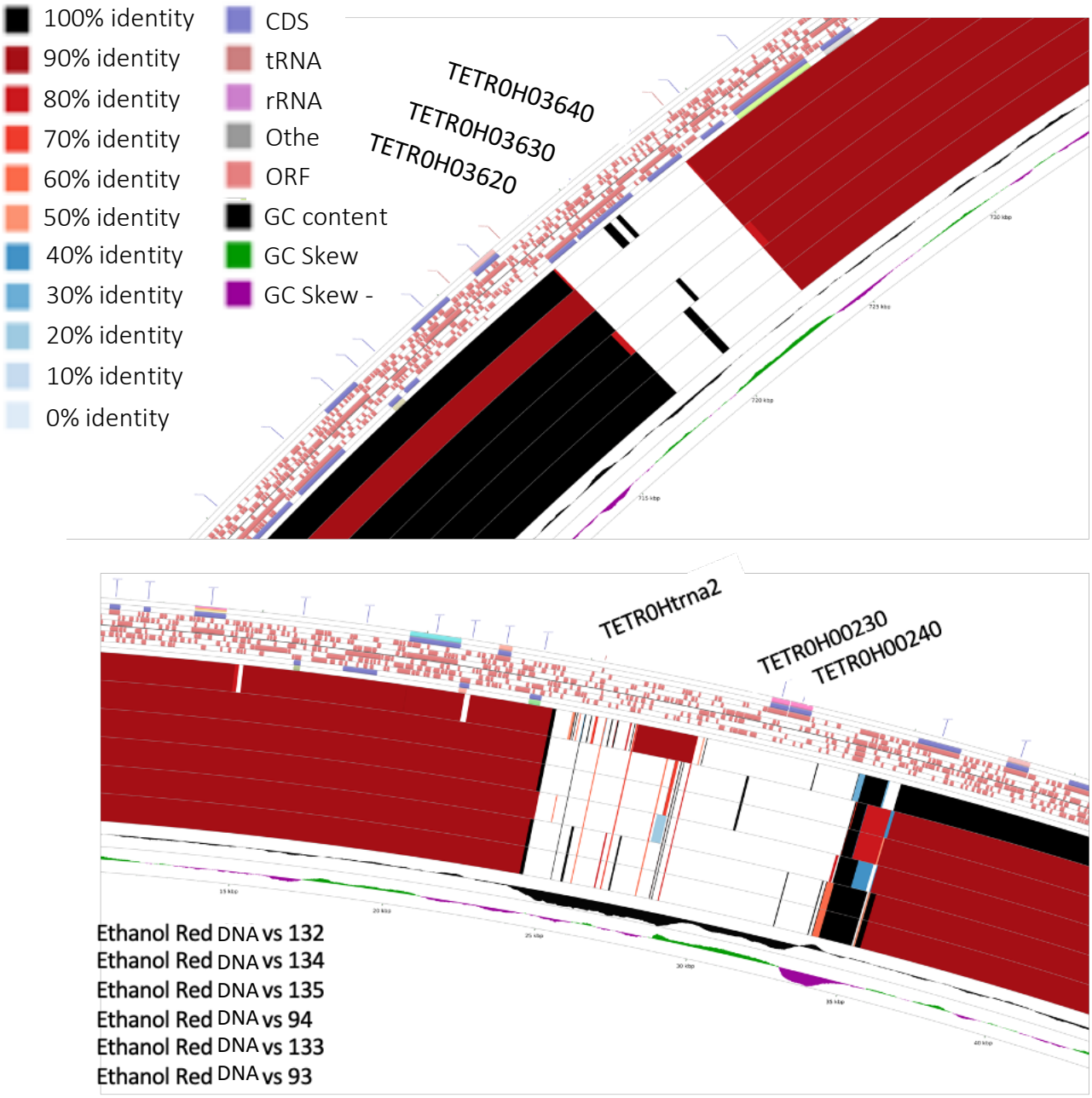


Figure 3.48 Zoomed CCT map of Chromosome 14. Generated using thecreate_zoomed_maps.sh command from the project used to create figures 3.14 and 3.15. A) is the 25,000 bp region zoomed by a factor of 12. B) is the 720,000 bp region zoomed by a factor of 12.

Chromosome 15, just over 1 mbp follows the pattern of having fewer predicted coding regions present in the telomeres of the lineage than Ethanol Red. This is shown in the CDS and DNA maps in figure 3.49 and 3.50 respectively, where it can be seen that the three most similar strains 135, 134 and 93 have high similarity on one telomere only while 132 and 133 have identity on neither telomere. Additionally, 133 has a large deletion of over 200 kbp at the start of the chromosome. Again, this can be attributed to a fault of the assembly rather than a genuine difference.

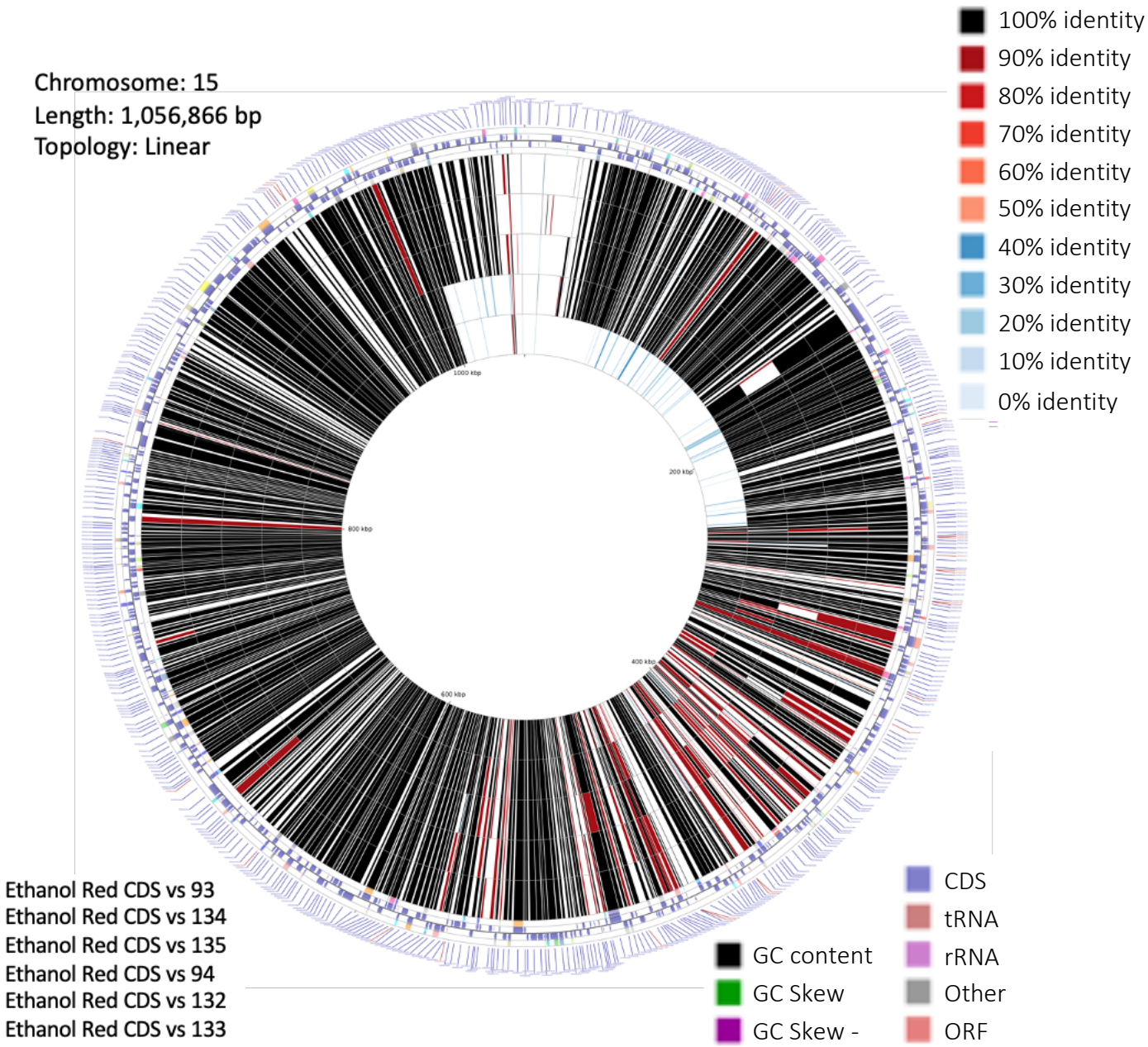


Figure 3.49 CCT map comparing the EtOH Red chromosome 15 with chromosome 15 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

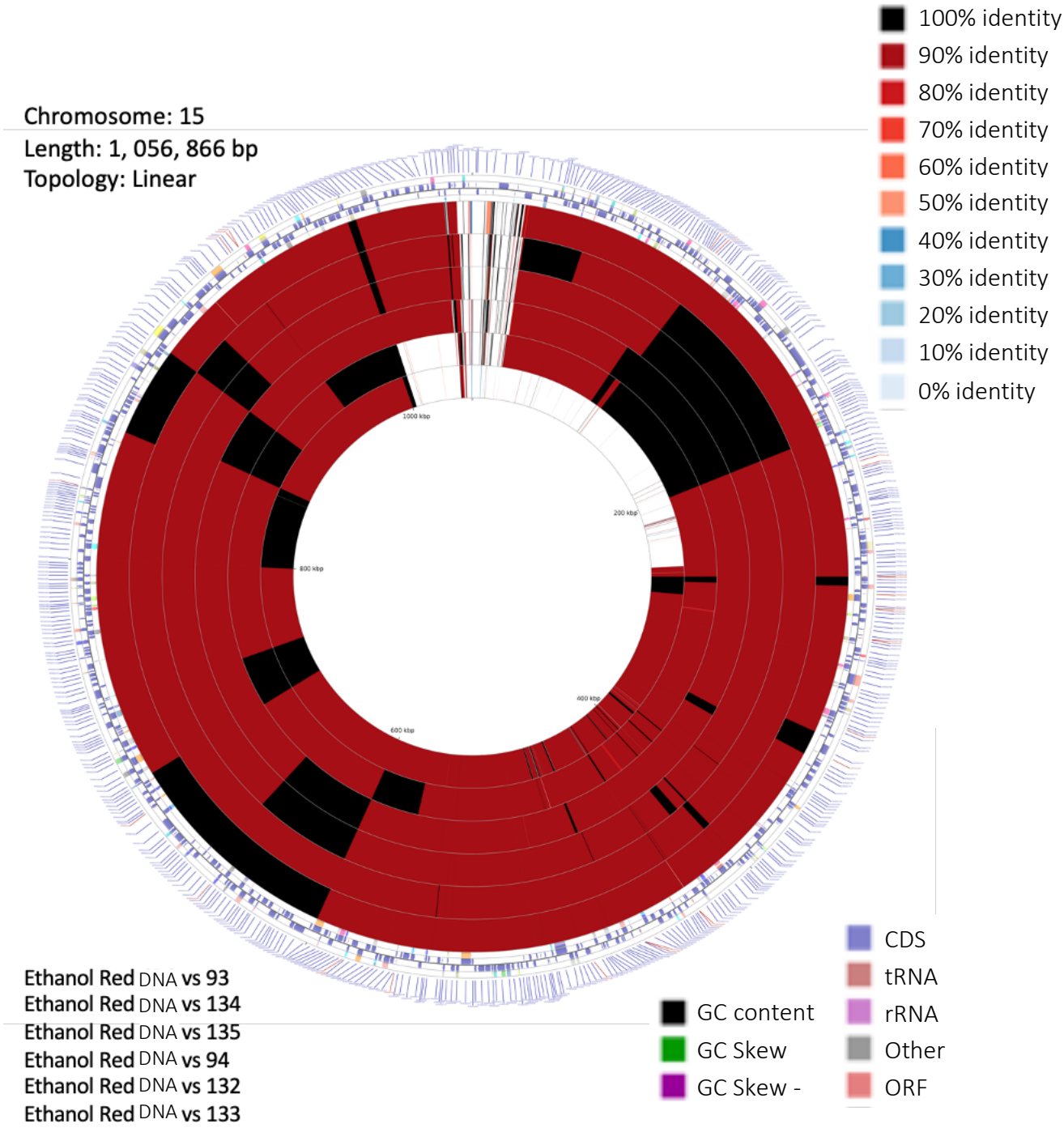


Figure 3.50 CCT map comparing the EtOH Red chromosome 15 with chromosome 15 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

Chromosome 16, with the CDS to CDS map shown in figure 3.51, has only one region where there is no identity and that is covering the first two coding regions on the chromosome. The first is unlikely to code for a function protein and the second, TETR0F00120, is FEX1 which is involved in fluoride export. The null phenotype makes a strain sensitive to fluoride, but as this has arisen as a result of the whole-genome duplication event the second copy, FEX2, means that a single knock out results only in a less marked fluoride sensitivity phenotype (Li *et al.*, 2013).

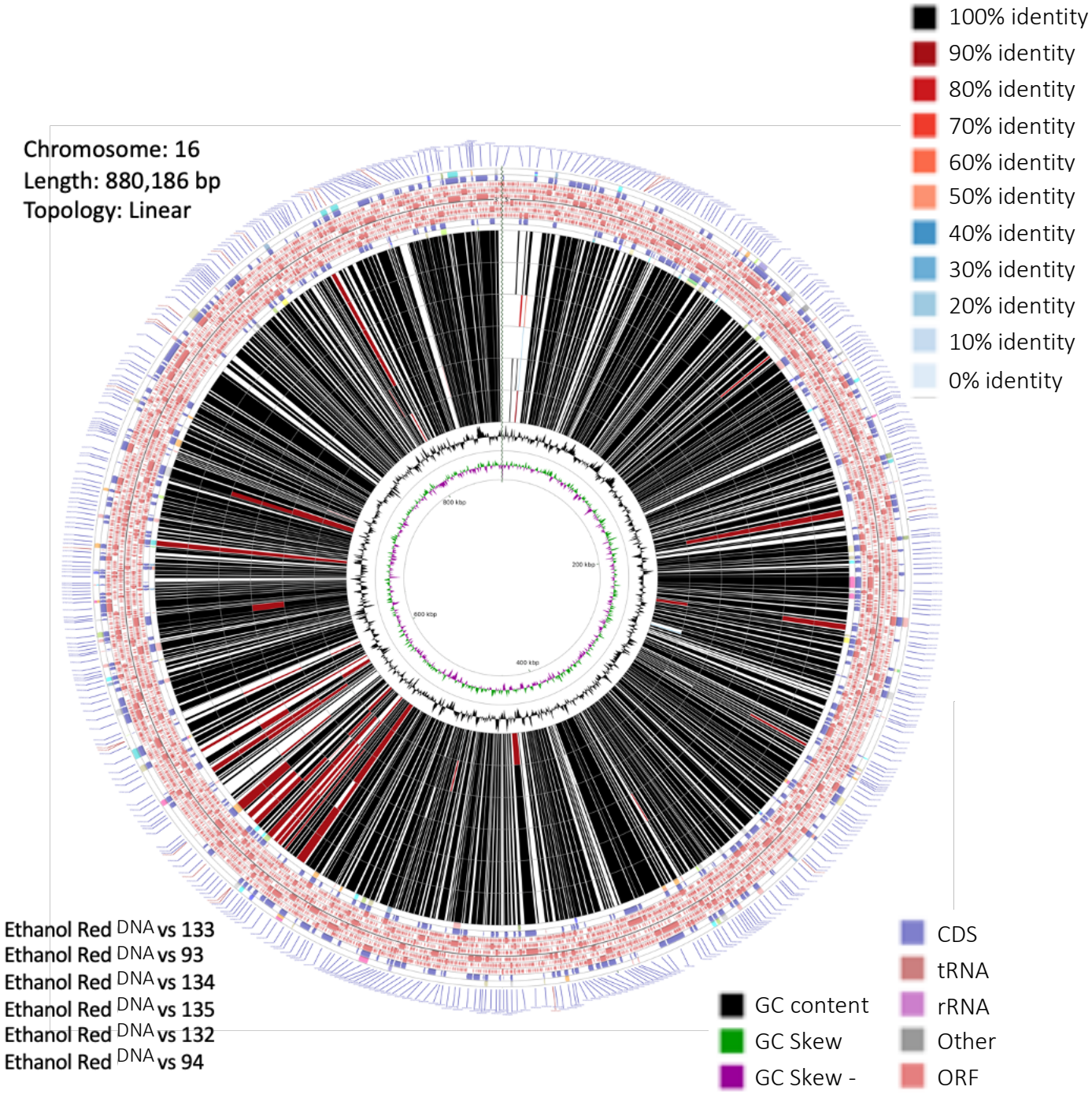


Figure 3.51 CCT map comparing the EtOH Red chromosome 16 with chromosome 16 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of CDS translations between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

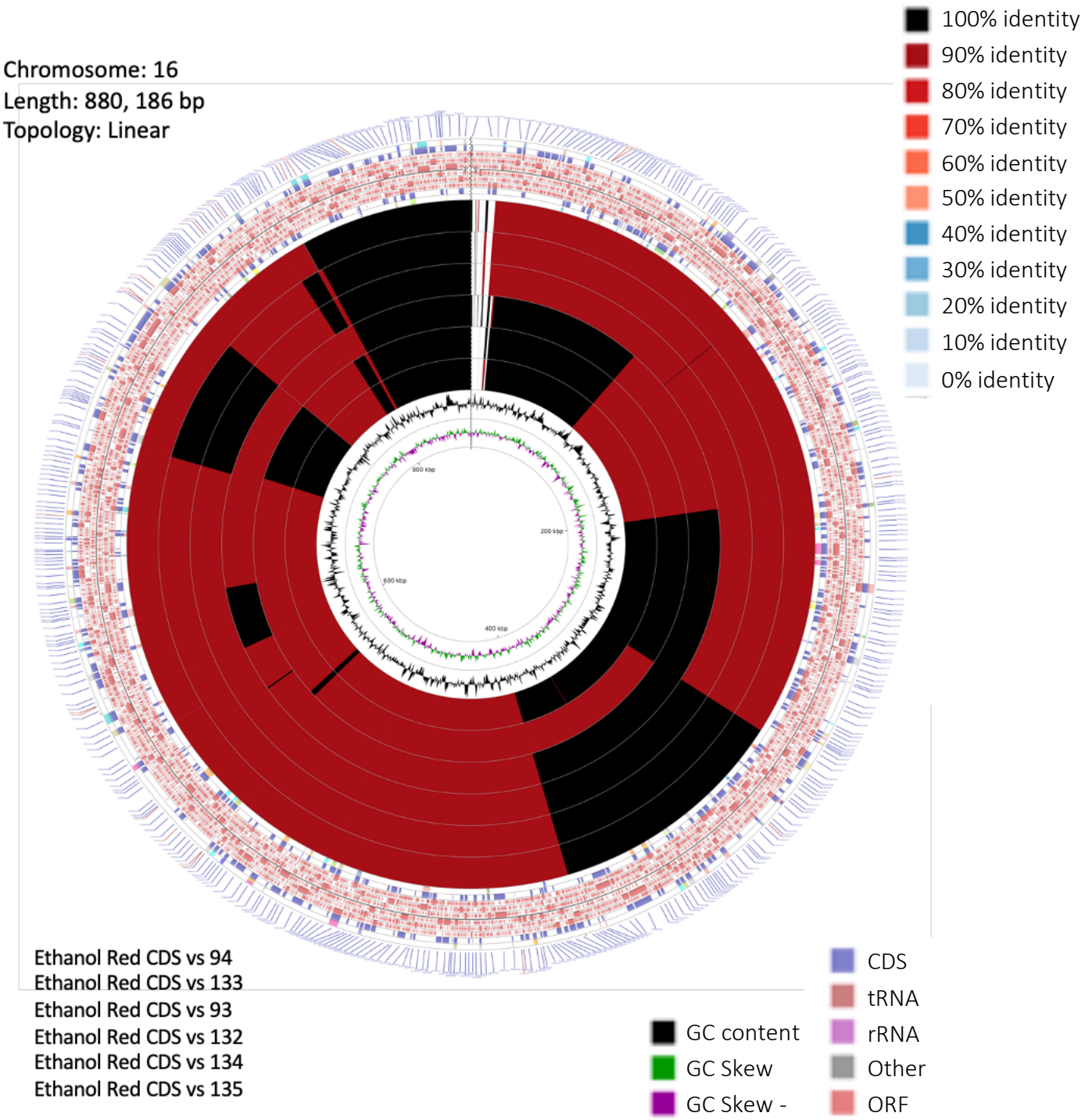


Figure 3.52 CCT map comparing the EtOH Red chromosome 16 with chromosome 16 from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

The mitochondrion has few coding sequences predicted but as can be seen in figure 3.53 there is a very large degree of sequence identity throughout all strains.

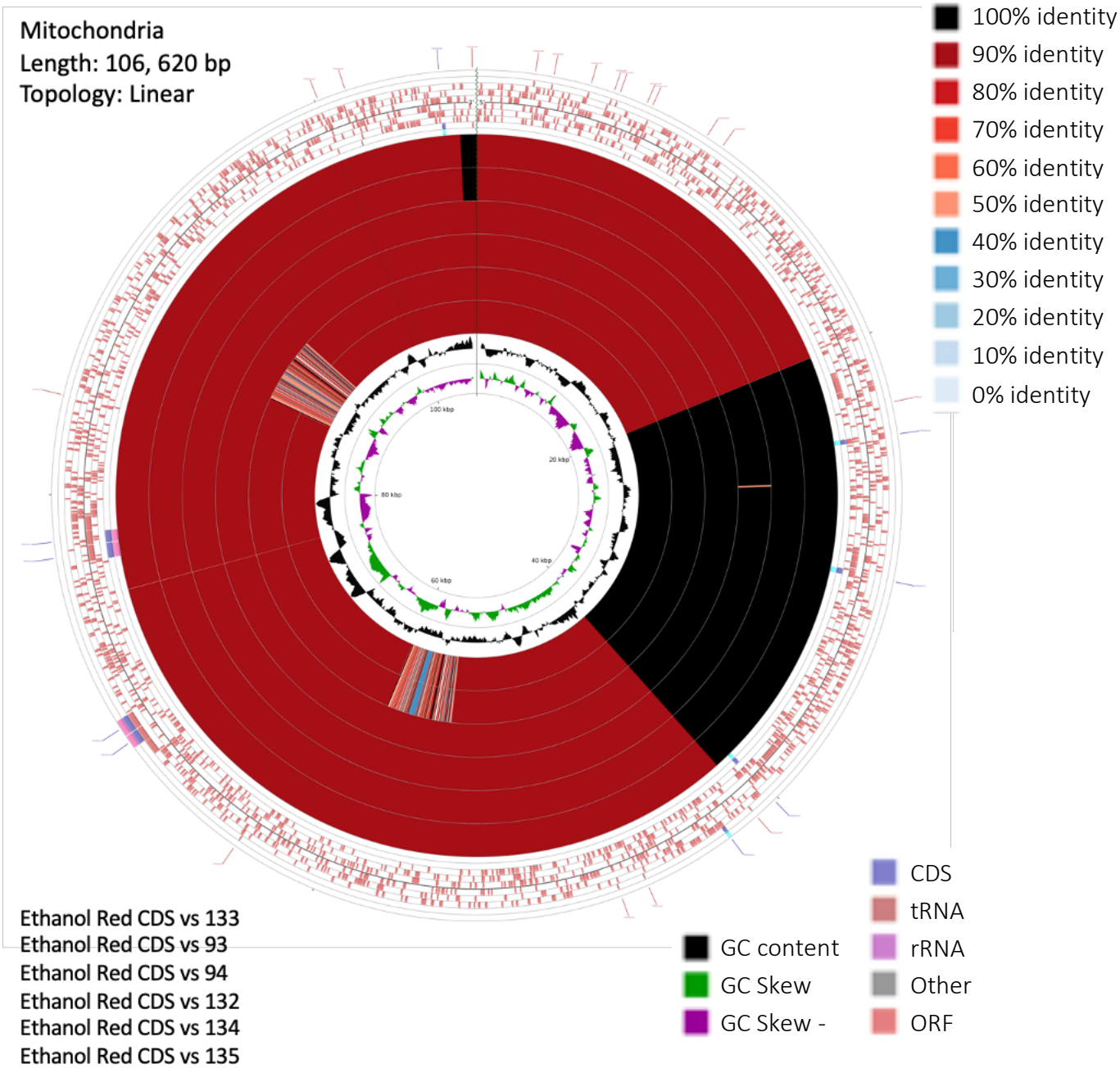


Figure 3.53 CCT map comparing the EtOH mitochondria with mitochondrial sequence from 6 other strains. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. COG functional categories for forward strand coding sequences; 2. Forward strand sequence features 3; Forward strand open reading frames; 4. Reverse strand open reading frames; 5. Reverse strand sequence features; 6. COG functional categories for reverse strand coding sequences. 7. the next 6 rings show regions of similarity as detected by BLAST comparison of DNA sequence between Ethanol Red and the 6 comparison genomes in the order indicated in the figure. 8. GC content 9. GC skew

3.2.4.2 Single Nucleotide Polymorphisms

Snippy was used to determine the location and type of single nucleotide polymorphism in the strains of the lineage compared to the Ethanol Red genome. The resulting SNPs found in the lineage that corresponded to predicted coding regions are given in supplementary tables 1.1-1.6. A graphical representation of the location of these SNPs in each of the lineage strains relative to Ethanol Red is given in Figure 3.54. Similar to the larger deletions seen in the CCT maps it is seen here that there are a large proportion of smaller genomic differences near to the telomere ends. It can also be seen here that there is a large convergence of mutations with very little variation. The two different branches of the lineage can also be determined here as with the mutations shown being stable, they continue through each lineage in two divergent patterns. It is also noted that the majority of SNP's shown here are in regions of the chromosome already highlighted from the CCT maps as having low identity at a CDS level. Anecdotally, strain 94 and those derived from it have been observed to be more promising than those in the other branch of the lineage. As such the differences that are common to all strains in the 94 branch but not in the 93 branch have been investigated further. On chromosome four there is a SNP detected in strains 94, 134 and 135, this has been found to be a frameshift mutation in a gene likely encoding for a ty like retrotransposon. In Chromosome eight there are two mutations that appear by the end of the lineage in the 94 branch The first of these is a frameshift mutation in a proposed acid phosphatase gene which could have an impact on gene regulation in these strains.

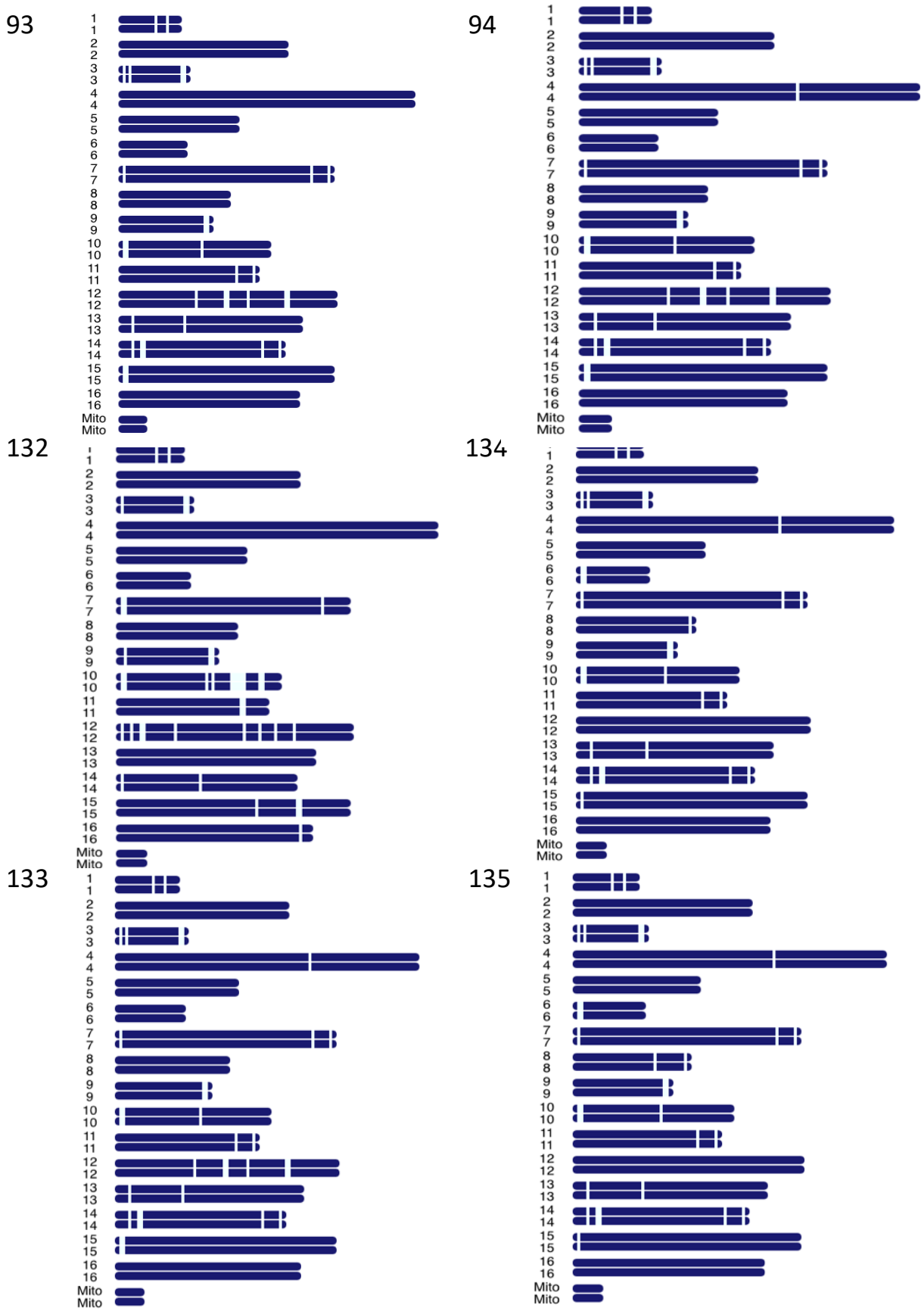


Figure 3.54 SNP Mapping. SNP locations determined by Snippy and plotted onto chromosomes of the reference genome Ethanol Red using chronomap in R. In each case the chromosomes of Ethanol Red are shown in dark blue with the locations of relative SNPs from the strain indicated in pale blue.

3.2.1 Heterologous Gene Insertion.

Two plasmids containing maltose utilisation genes were inserted into the lineage yeast strains. PCR was used to test for the correct insertion of the MAL genes onto the chromosome. Figure 5.55 shows the resulting gel electrophoresis where a band of the expected size is observed in the transformed strains and not their wild-type counterpart. The primers annealing over the expecting insertion site yielded no positive bands. This may suggest that the insert was not successful. Further to this the curing plasmid failed to remove the resistance marker. It was shown that the curing plasmid was inserted as the resistance phenotype of that plasmid was conferred only to transformed strains and not the wild type. However, the phenotype of cured strains could never be achieved. A number of primer pairs were trialled and failed to confirm insertion at the expected site. The option to select for successful marker-less transformants by growth on maltose as would normally be utilised with this type of protocol was not available here as it was shown in early phenotypic testing of these strains that they all grow equally well on maltose as a sole carbon source as is shown in figure 5.56. It is proposed that the reason for continuous failure of cloning is due to the homologous arms that were designed for the specificity of the integrative vector were not optimal as they were designed before the complete sequence was obtained.

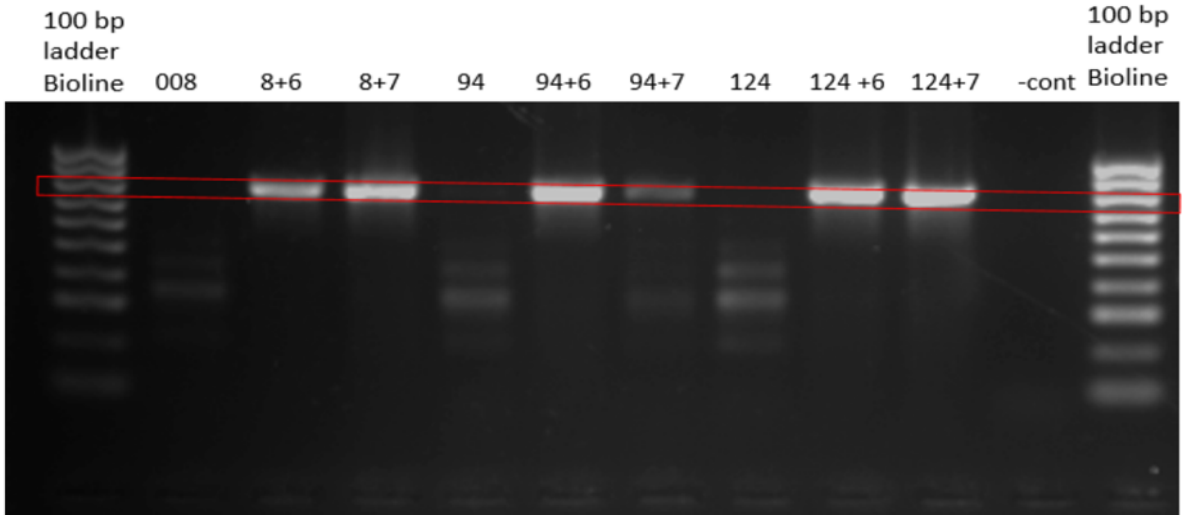


Figure 3.55. Gel electrophoresis of genomic DNA from wild type and transformed strains.

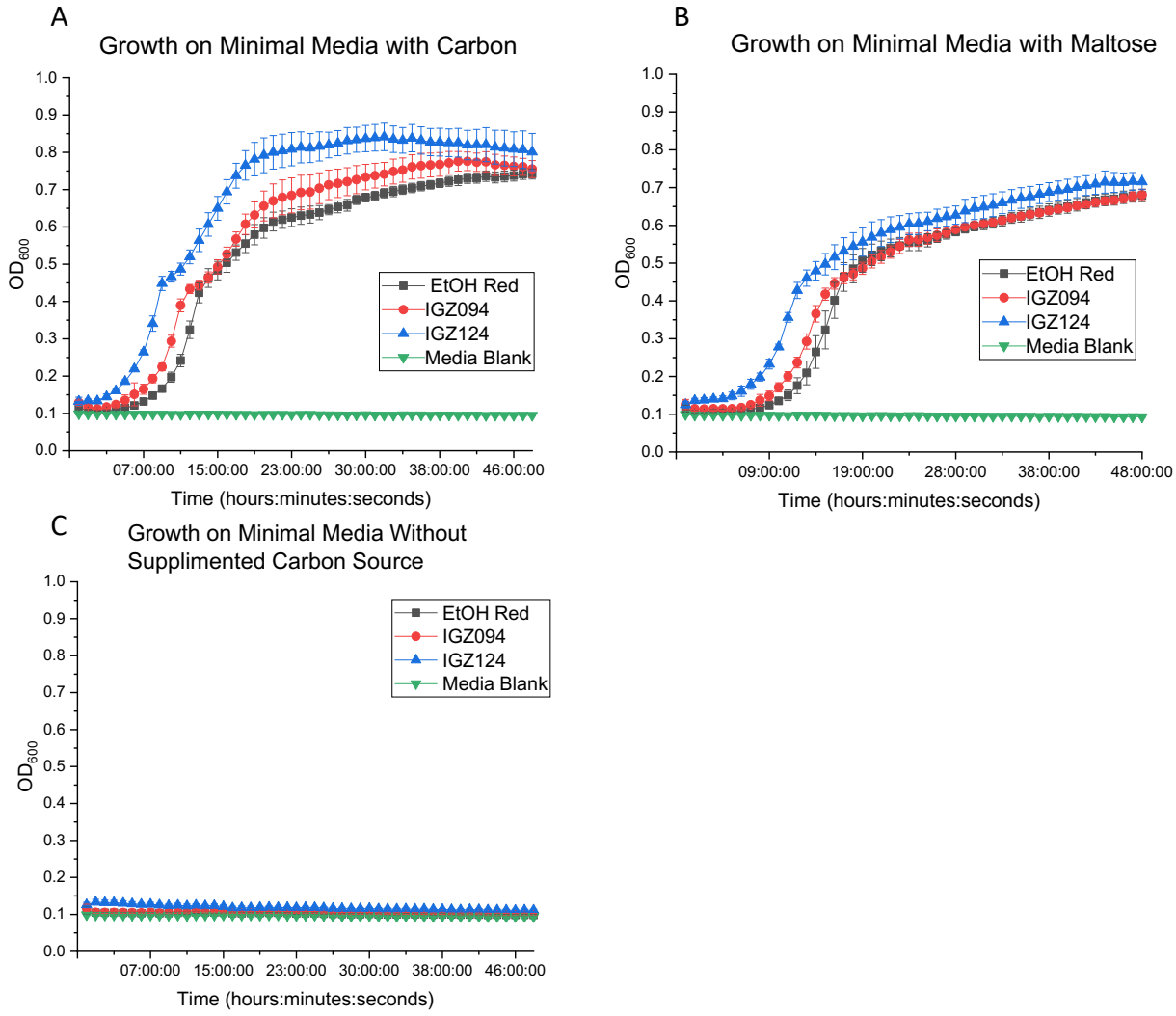


Figure 3.56. Growth on maltose as a sole carbon source. The three "wild type" strains that were the recipients for the inserted maltose genes. A) growing on carbon as a sole carbon source. B) growing on maltose as a sole carbon source. C) In minimal media without carbon source as a negative control. In each case grey is Ethanol Red, 94 is red, 124 is blue, and the media is green.

3.3 Conclusions

The trade-off here allows for long read sequencing to only be performed on a fraction of the strains vs the lack of telomere information is thought to be acceptable for the purposes for which our industrial partners wanted this data, but it may prove limiting in other cases.

The use of long read sequencing with short read sequencing for correction and the subsequent use of the resulting genome as a reference has been successful here, as the assemblies have proven to be informative. There are however several issues identified with these assemblies including the lack of successful sequencing of telomeric regions and the apparent misassembly of several genomic regions throughout the lineage. Ideel assessment of these assemblies showed that whilst there were misassembled genes, there was also a marked improvement between the long read only assembly and the Illumina read corrected version. This could potentially be improved upon by further alternations to the sequencing data analysis and assembly pipeline, though these improvements are likely to only be marginal. Improved assemblies could be obtained by further long read sequencing by PacBio or more Nanopore reads to give improved sequencing depth and supplementing with Sanger sequencing over problematic regions and for confirmation.

Despite the issues discussed above; these sequences provided insight that was previously not clear from the published data. The improved yeast sequences here were able to answer several questions pertaining to the industrialistion of these strains. One particularly important confirmation came from the specific use of long read sequencing chemistry. The insertion of single copy genes in these strains was previously confirmed by qPCR for which, while this is a robust mechanism, long read sequencing can complement this information. The presence of a gene can be shown by sequencing over the entirety of the gene and the flanking regions in one single long read as well as the opposite situation where only the flanking regions are seen (this situation is illustrated in figure 5.57). Data of this type, when provided with high coverage is particularly useful as it excludes the possibility of experimental artefacts.

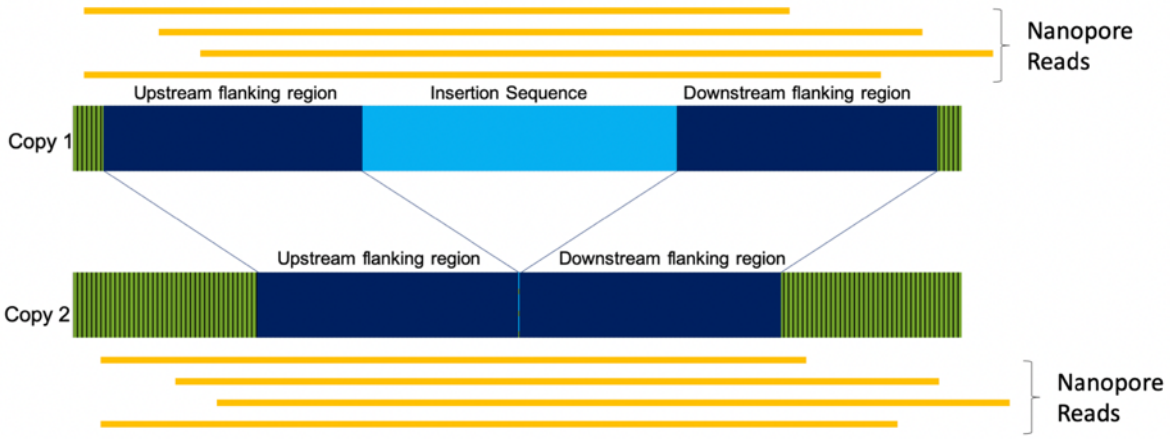


Figure 3.57. Illustration of long read sequencing in diploid organisms. Yellow lines depict individual raw nanopore reads, pre-assembly. In light blue is an inserted sequence on one chromosome copy.

Chapter 4. Relating Phenotype to Genotype: Omnilog assays

4.1 Introduction

The results obtained in chapter three provided a good basis for understanding the genomic capabilities of the *S. cerevisiae* strains used in this study. The next step was to investigate how the genomic differences observed are manifested as a phenotype multiple experiments were carried out at a variety of scales. The three strains focused on in this chapter will once again be the industrial strains Angel and Ethanol Red as well as the lab reference strain S288c. Initially a bioreactor fermentation was carried out to understand how the strains in question differ, in a rich medium where the growth conditions are more comparable to that used in industry than they are in microlitre scale on a plate. To obtain a full panel of growth comparisons over different conditions the Omnilog Biolog was used. As is outlined in the main body of the introduction the Omnilog Biolog is a high-throughput phenotypic microarray system that measures cellular respiration as a proxy for growth over time in plates that contain a variety of different growth conditions including carbon and nitrogen sources, pH and osmolytes. The Biolog phenotype arrays were run with multiple strains in parallel in order to compare the growth of the strain in the various conditions and use this information to corroborate the results from the genomic analysis. In addition to this high throughput, general approach these results will be used to inform further growth experiments with conditions of interest to authenticate the Biolog results.

4.1.1 Aims and Objectives

This chapter aims to obtain and analyse an array of phenotypic data for the strains of interest as well as the lab reference strain S288c. This data will be analysed to determine the organism with the ability to utilise the most carbon sources and improved stress response. Finally this chapter aims to relate this data to the genotypic data already obtained and provide an analysis method for Biolog data that meets the needs of ease of use and good data visualisation.

4.2 Results

4.2.1 Establishing Fermentation conditions in a Parallel Bioreactor System

To test the strains in a small-scale fermentation the DASGIP parallel fermentation system was used with 5L cultures of a rich medium to allow for comparison of the strains when grown simultaneously in the same conditions. A culture of pre-inoculum of the lab strain S288c, unsuitable for fermentation and monitoring by OD600 due to flocculation, which is caused by the expression of several *FLO* genes. This is often a preferred phenotype in the brewing industry as it facilitates easy separation of the cells from media and products (Soares, 2011). Though in this case it would have made monitoring of the fermentation difficult, specifically it would lead to inaccurate OD measurements in this system. The industrial strains, Angel and Ethanol Red were each grown in duplicate with a maintained temperature of $30^{O}C$ and the CO_2 production was monitored continuously for the duration of the fermentation (27 hours). The CO_2 production over time is given in figure 4.1 and shows no difference in any measurement during the fermentation including ethanol production, which is shown by CO_2 output in *figure 2*. CO_2 production is directly proportional to ethanol production as the breakdown of one molecule of 6-carbon glucose yields two CO_2 molecules and one molecule of ethanol. This

fermentation established that there is little measurable difference between these two strains, in terms of ethanol production in rich medium. As such, deeper strain investigation is required to elucidate the differences between these strains.

From the parallel fermentation, it can be seen that at this level there is no significant difference between the two different industrial strains. This data offers a good benchmark for the rest of the study as it shows that the basal functionality of the two different strains is very similar in terms of growth rate and ethanol production under standard glucose fermentation conditions. However, the reason for testing different *S. cerevisiae* strains for ethanol production is to find the strain that can cope best in sub-optimal conditions potentially in the presence of inhibitory molecules. Overall, this initial experiment allowed for the creation of a standard method for use and a standard for comparison. It would be interesting to repeat this experiment with the inclusion of S288c as a control.

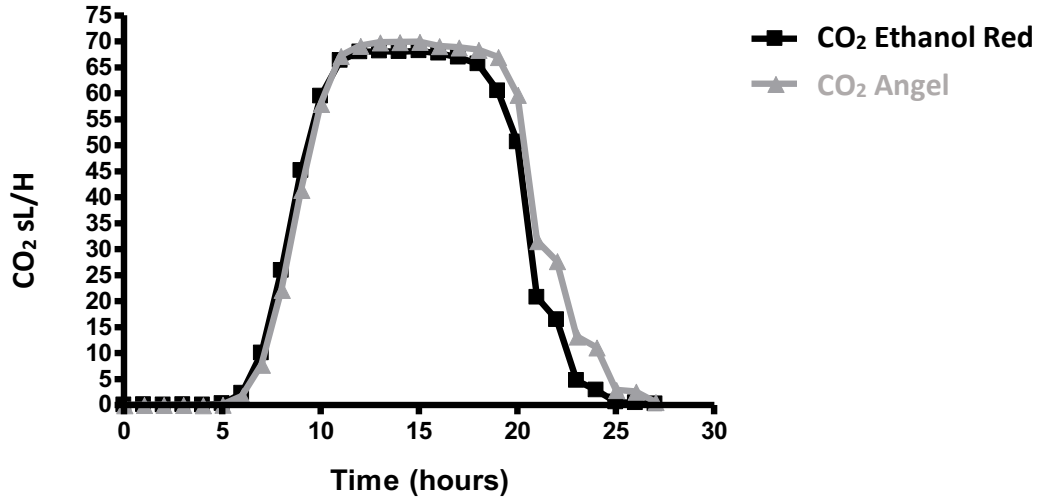


Figure 4.1 CO_2 production during fermentation. The CO_2 produced in standard litres per hour by both Ethanol Red and Angel over the course of 27 hours. P = 0.0082 using a two- way ANOVA test.

4.2.2 Biolog Phenotypic Arrays

4.2.2.1 Omnilog Results

Biolog phenotypic microarrays were used to assess the growth of S288c, Ethanol Red and Angel over a vast range of growth conditions simultaneously. The heat maps generated from the Omnilog output show that there are three distinct types of results from the comparison: when both strains grow in a similar fashion i.e are equally able to grow in a particular condition, when strain one (or the reference strain in the context of the analysis) grows better than strain two, or the opposite when strain two grows better than strain one. The three yeast strains, S288c, Ethanol Red and Angel were grown in 6 different Biolog PM plates; PM1 and PM2 with carbon sources, PM3 Nitrogen, PM5 Nutrient Supplements, PM9 Osmolytes and PM10 pH. This phenotypic data was considered together with the genomic information for each strain to corroborate any differences in growth phenotype with a difference in genomics that leads to changes in the metabolic pathways for utilisation of the metabolites etc in the PM plates. The data visualisation used in this chapter shows the result of two growth curves being compared as whether or not the strain has improved growth compared to the reference is more valuable information than simply whether or not the strain grows. In each graph the dendrogram groups conditions in which similar growth is seen. In all cases, a more detailed version of the results is given in supplementary figures 2.1-2.6.

4.2.2.1.1 PM3 Nitrogen

Compared to the first plates for carbon utilisation more drastic differences can be seen in the nitrogen utilisation plates, as such we will begin by analysing PM3 Nitrogen. As can be seen in figure 4.2 when comparing the adapted industrial strain Ethanol Red with the lab strain S288c both strains begin growing in a similar manner in almost all conditions until t = 4.5 hours where Ethanol Red begins to grow more in certain conditions. A distinct improved growth pattern can be seen with 38 out of 96 of the conditions, these are all grouped on the left of the graph by the dendrogram. There are no instances on this graph where the same level of growth difference is seen in the opposite situation where S288c outperforms Ethanol Red. This is to be expected since a large directive in the improvement of *S. cerevisiae* for industrial ethanol production is a widening of its nutrient utilisation profile. The majority of the graph, shown in pink, is where both strains grow at a similar level in that particular condition. A similar pattern is seen when Angel is compared to S288c on this plate as is shown on figure 4.3. Since both strains have been adapted for the purpose of ethanol production it is possible that they would both have been developed independently to utilise similar growth mediums and thus now have similar nutrient utilisation profiles. To further show the similarities between Angel and Ethanol Red they were compared to each other, this is shown in figure 4.4. Here it can be seen that in almost all of the conditions both strains grow to a very similar level. Despite one condition, α-Amino-N-Valeric Acid, where in the mid-point of the growth period Angel has slightly more growth by the end point both Ethanol Red and Angel have balanced to the same level of growth. What isn't clear from looking at both strains compared to the reference S288c is that overall, Ethanol Red is slightly 'fitter' in terms of growth on diverse nitrogen sources compared to Angel. In several conditions where yellow can be seen indicating Ethanol Red exhibiting more growth than Angel, it appears more in mid growth phase before evening out towards the end point of the experiment. This is seen most prominently in H4 Ala-Gly and H7 Ala-Thr and

indicates that Ethanol Red can grow faster on these nitrogen sources but that both strains still grow until some other limiting factor prevents further growth. There are 10 conditions in which Ethanol Red exhibits improved growth over Angel these are, Homoserine, Gly. Glu, Met.Ala, Tryptophan, δ-Amino-N- Valeric Acid, Ala.Asp, Threonine, Proline, citrulline, biuret and most notably Glycine.

Genomically, it has been found that S288c doesn't have the GCV1 gene which encodes for part of glycine decarboxylase which is required for glycine breakdown (JB *et al.*, 1997). Nor does it have CHA1 which would allow for threonine metabolism (JG, Petersen *et al.*, 1988). More generally, the improved nitrogen source utilisation observed in Angel compared to Ethanol Red is likely due to the three oligopeptide transporters found in Angel but not Ethanol Red. The group of fungal oligo peptide transporters have been found to provide an advantage in the context of wine yeast and it is thought that these genes were acquired by horizonal gene transfer from *Torulaspora microellipsoides* (Becerra-Rodríguez, Marsit and Galeote, 2020).

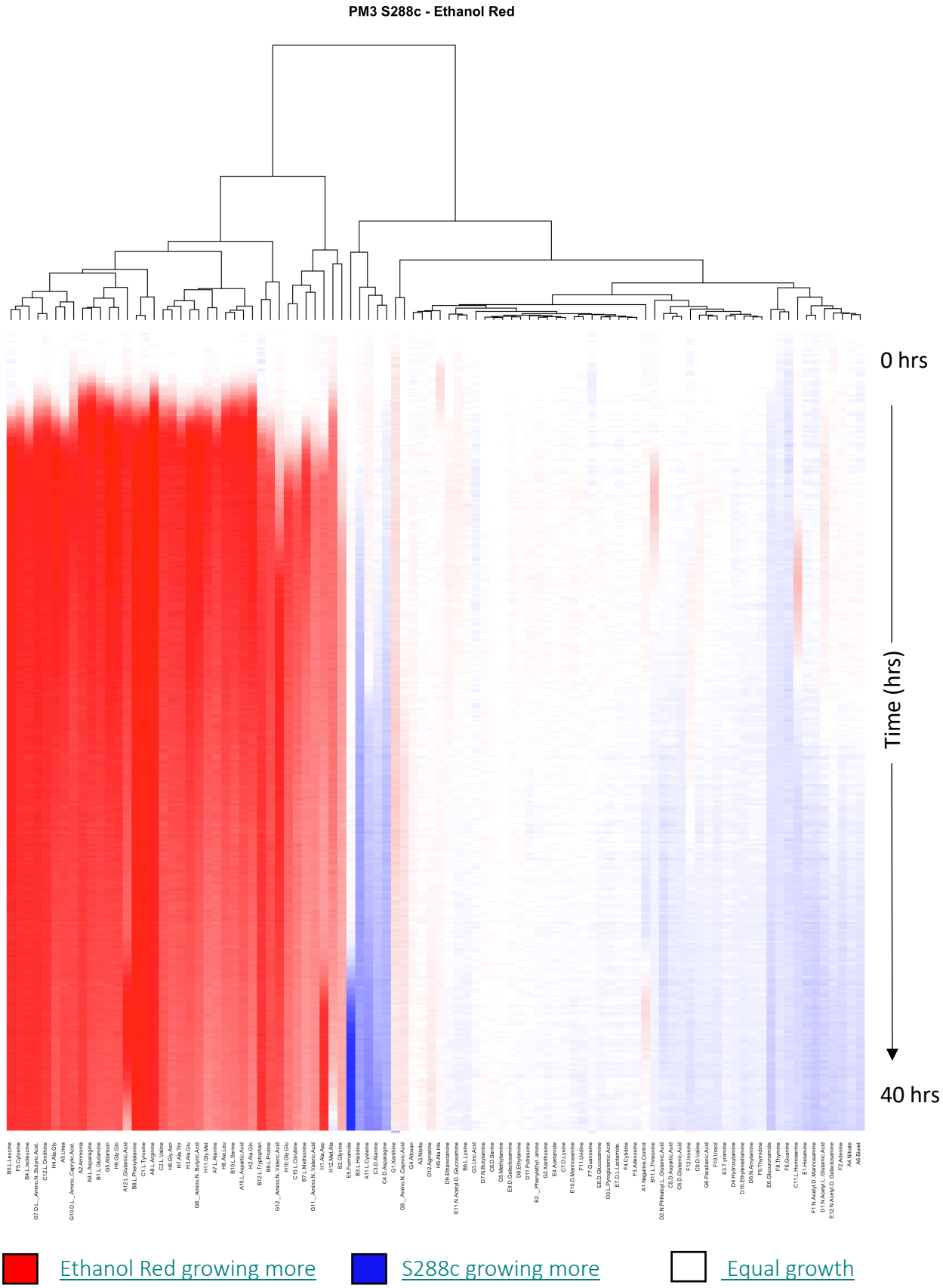


Figure 4.2. Heat map showing the growth of S288c and Ethanol Red on PM3 nitrogen. Growth time increases down the y axis. With each column representing a different well on the plate with a different nitrogen source. Red depicts time points/conditions where Ethanol Red is respiring more than S288c. white areas depicts time points/conditions where both strains are growing equally well. Blue depicts time points/conditions where S288c is respiring more than Ethanol Red. The dendrogram shows the grouping of similar relative growth rates.

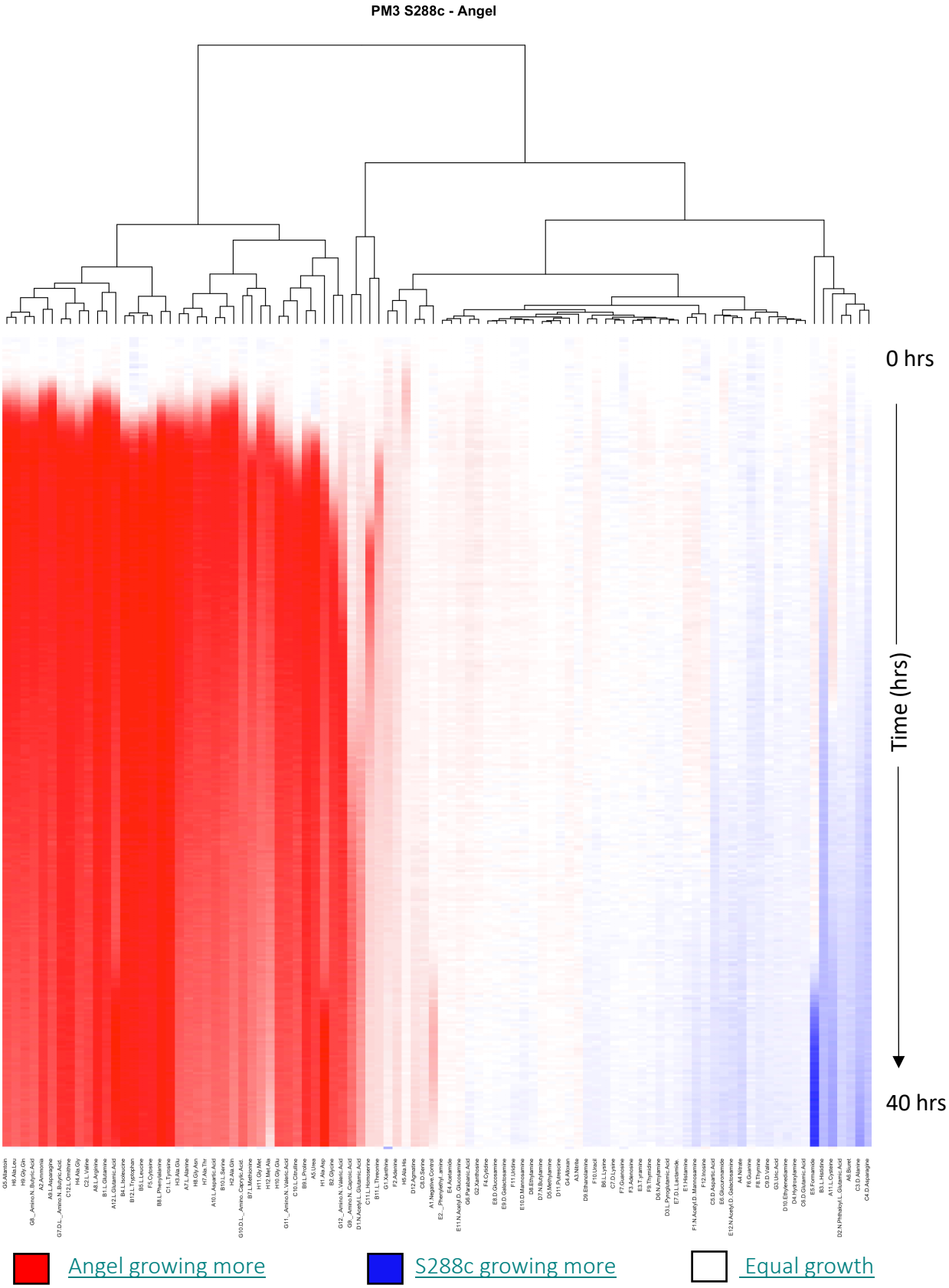


Figure 4.3. Heat map showing the growth of S288c and Angel on PM3 nitrogen. Growth time increases down the y axis. With each column representing a different well on the plate with a different nitrogen source. Red depicts time points/conditions where Angel is respiring more than S288c. White areas depicts time points/conditions where both strains are growing equally well. Blue depicts time points/conditions where S288c is respiring more than Angel. The dendrogram shows the grouping of similar relative growth rates.

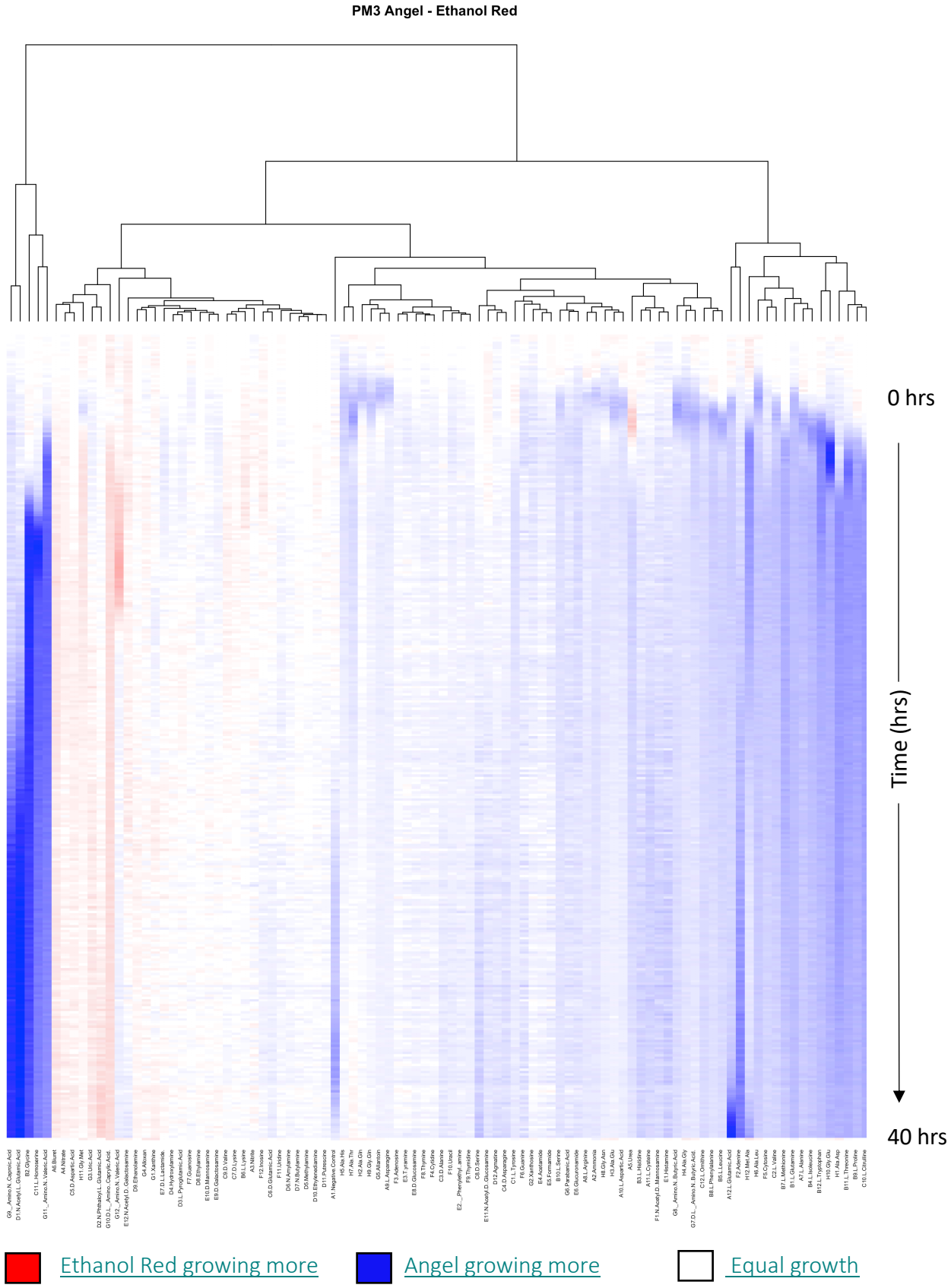


Figure 4.4. Heat map showing the growth of Ethanol Red and Angel on PM3 nitrogen. Growth time increases down the y axis. With each column representing a different well on the plate with a different nitrogen source. Blue depicts time points/conditions where Angel is respiring more than Ethanol Red. white areas depicts time points/conditions where both strains are growing equally well. Red depicts time points/conditions where Ethanol Red is respiring more than Angel. The dendrogram shows the grouping of similar growth rates.

4.2.2.1.2 PM1 Carbon plate 1

The plates for carbon course utilisation plates show fewer marked differences overall. Both Ethanol Red and Angel when compared to S288c have at least slightly improved growth on most carbon sources, which is shown by a faint purple colour in most of the columns. There are also multiple conditions where either Ethanol Red or Angel outperform S288c as can be seen in figure 4.4 and 4.5. Both industrial strains have improved growth on galactose, maltose, sucrose, and acetoacetic acid compared to S288c. Additionally Ethanol Red also has improved growth on saccharic acid and xylose.

To explain the improved growth on galactose the gene GAL1, which converts galactose to glucose, was searched for in all strains (Ramos, Szkutnicka and VP, 1989). Ethanol Red was found to have two copies of this gene while S288c and Angel had only one. This doesn't explain the growth difference between Angel and S288c. GAL7 and GAL10 were all annotated in all three genomes. Further, all strains are annotated as having a copy of the GAL80 gene as well as GAL4 and GAL3, the regulators for this pathway (Escalante-Chong *et al.*, 2015). At the SNP level, both industrial strains have several mutations in the GAL1 gene of which four are non-synonymous and present at the same points on both strains. No SNPs were found in either GAL80 or GAL4. Whilst all strains contain at least one copy each of MAL11, MAL12 and MAL13 which should confer the ability to utilise maltose, S288c cannot. S288c is known not to have the required regulatory genes to transport and utilise maltose (Duval *et al.*, 2010). Sucrose metabolism requires the gene SUC2 which is present in Angel and Ethanol Red but absent in S288c (Carlson, Osmond and Botstein, 1981). Two genes have been shown to be involved in the degradation of acetoacetic acid in yeast, ACS1 and ACS2. Whilst ACS1 can be found in all strains, ACS2 is found only in the industrial strains which may explain the differing phenotype. Looking at the genomic data it can be seen that Ethanol Red has XylE a D-xylose proton-symporter required for xylose utilisation that were not found by BLAST search in the S288c genome (Richard, MH and Penttilä, 1999). Which may explain the improved growth on this carbon source for Ethanol Red, however, Angel also has this same phenotype but does not have the XylE gene. There is some redundancy in this system though and Angel has Xyl2 which is also a xylose transporter. This would explain all three phenotypes, but not quite, as S288c also has this second transporter. Since it was observed that S288c does grow, to a lesser extent than the two industrial strains on this medium it can be hypothesised that the difference here is due to an ability of the industrial strains to grow to a higher density than S288c even in favourable conditions. To this end the comparison to the growth on glucose, the preferred carbon source for *S. cerevisiae* shows a very similar pattern.

In both first set of carbon utilisation plates there is no condition where S288c shows improved growth over either of the industrial strains. When Ethanol Red and Angel are compared to each other there is only one case where there is a very clear instance of a greatly improved growth of Ethanol Red over Angel, G10 methyl pyruvate, though a majority of the conditions show at least marginally more growth of Ethanol Red compared to Angel. On pyruvic acid and acetoacetic acid Angel has improved growth.

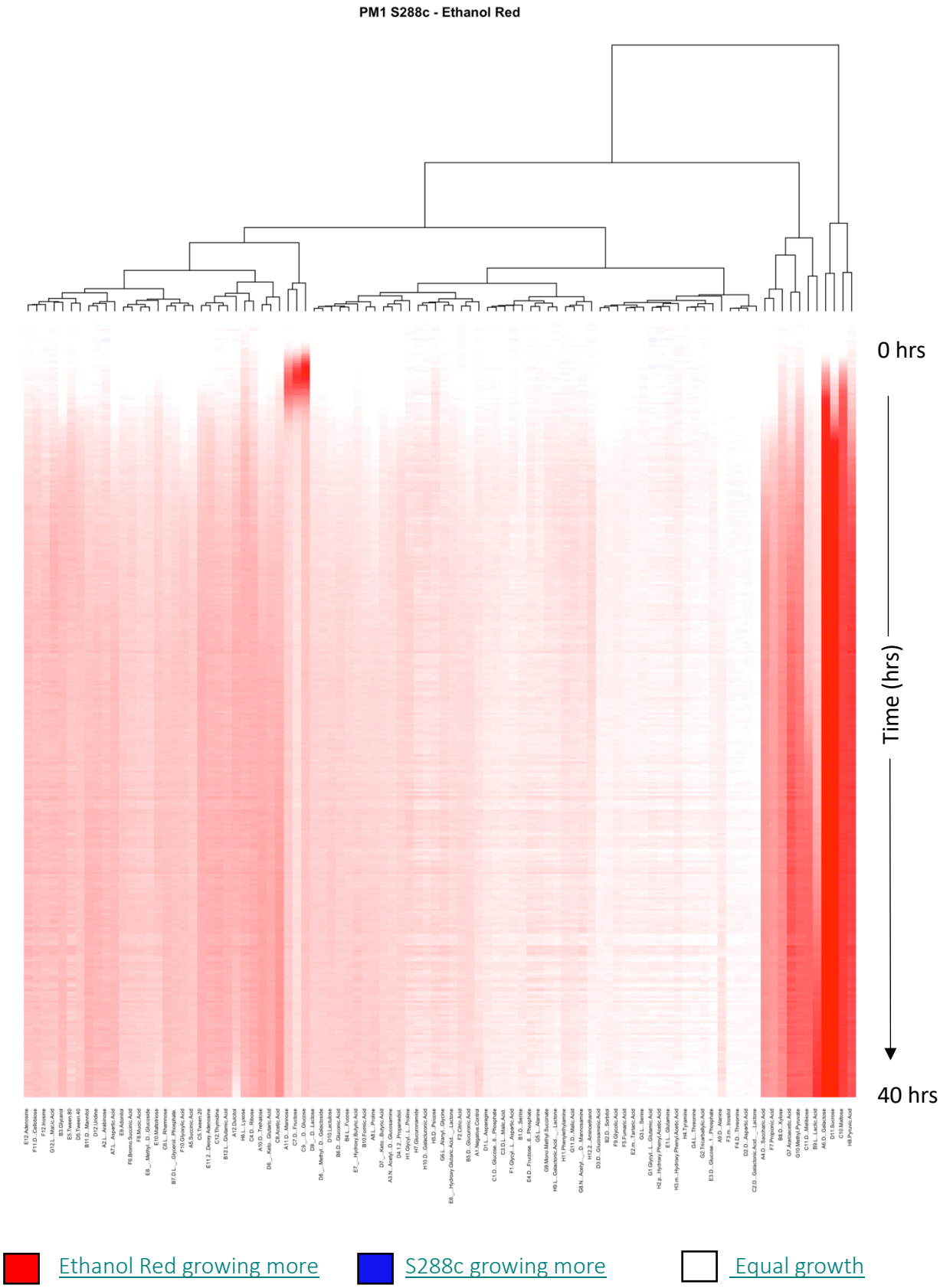


Figure 4.5. Heat map showing the growth of Ethanol Red and S288c on PM1 carbon. All parameters and colours scheme the same as figure 4.2.

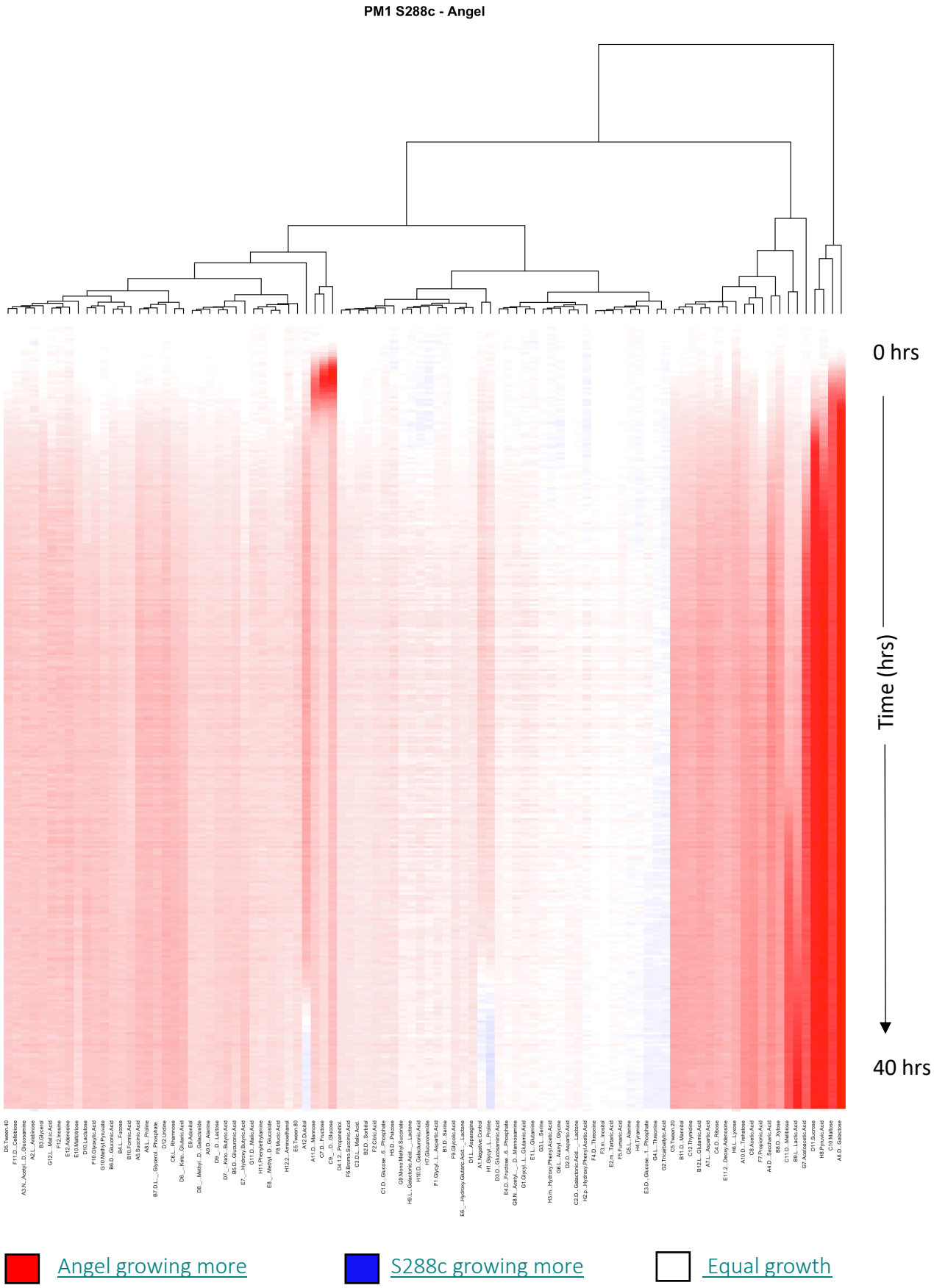


Figure 4.6. Heat map showing the growth of Angel and S288c on PM1 carbon. All parameters and colours scheme the same as figure 4.3.

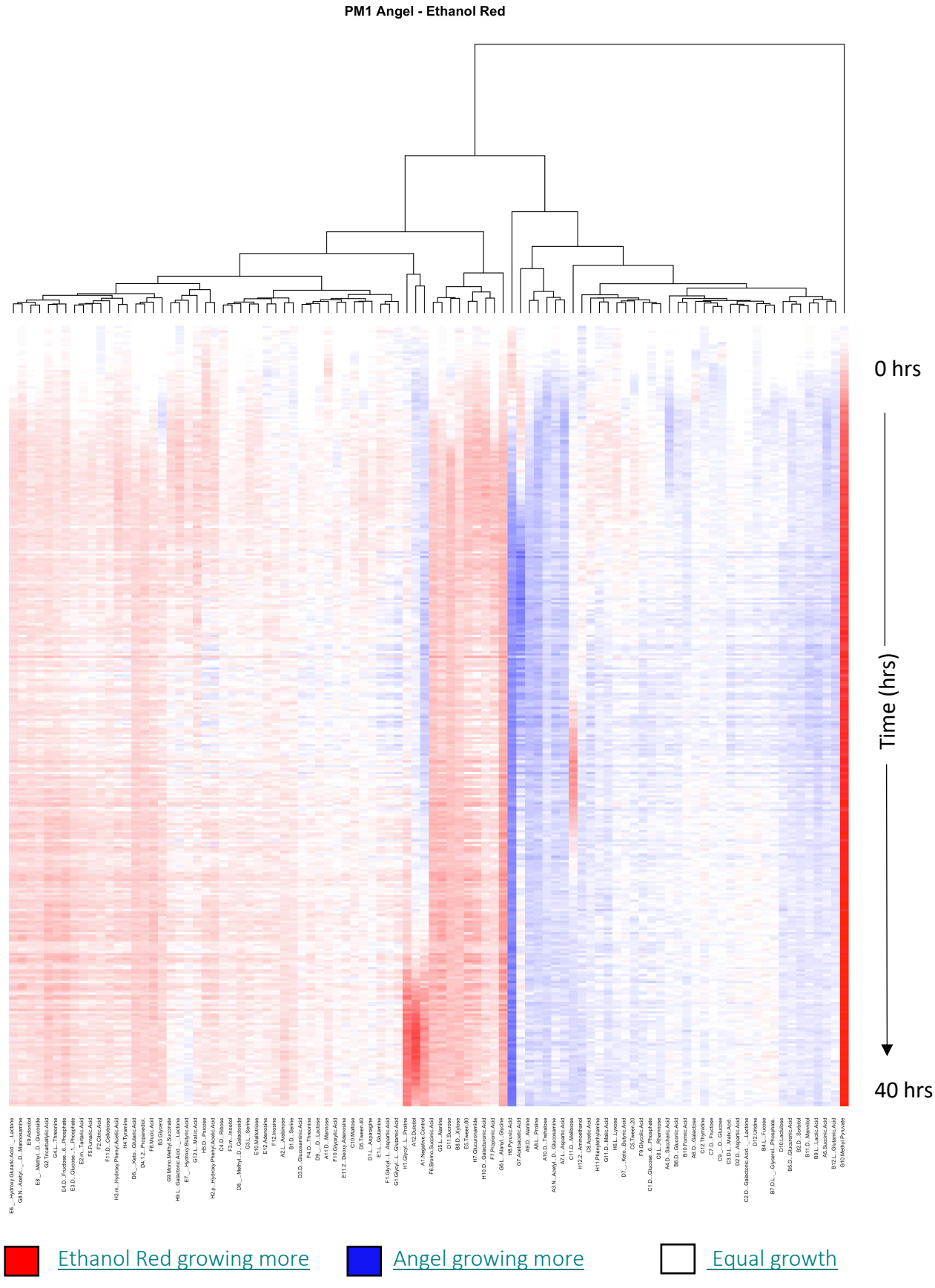


Figure 4.7. Heat map showing the growth of Ethanol Red and Angel on PM1 carbon. All parameters and colours scheme the same as figure 4.4**.**

4.2.2.1.3 PM2 Carbon plate 2

The second carbon plate shows very similar results to the first with S288c outperforming the industrial strains in no condition, and the two industrial strains showing greatly improved growth in only one condition with slightly improved growth in several other conditions. When the two industrial strains are compared to each other in this way there is more variation between strains than in PM1 but with each strain having improved growth in a similar number of conditions.

Ethanol Red shows much better growth on turanose compared to S288c and slightly improved growth in inulin, raffinose and methyl D galactoside. When compared to Angel in these conditions Ethanol Red has considerably better growth in methyl D galactoside and slightly improved growth in stachyose, hydroxybenzoic acid and arabitol.

Angel also shows improved growth on turanose and also on methyl D galactoside and raffinose compared to S288c. There are also smaller differences in growth on hydroxy benzoic acid, gelatine, arabitol, stachyose and pectin. When Angel and Ethanol Red growth rates are compared Angel has improved growth on Raffinose, acetamide and methyl D galactoside.
Genomic data would suggest that Ethanol Red is capable of producing histidine while Angel is not as it has histidinol dehydrogenase, Phosphoribosyl-AMP cyclohydrolase and Phosphoribosyl-ATP pyrophosphatase which are all involved in histidine biosynthesis and metabolism, but Angel does not (Smith and Ames, 1965). However, the phenotypic data shows that neither grows on histidine. Growth on stachyose and arabinose, appear to have differential growth patterns between strains though auto-reduction of the dye in these cases is possible and indicated by the shape of the growth curve, this was corrected for by normalising the data by subtracting the time zero value from all other time points for each well.

Turanose can be accepted into the maltose utilisation pathway at a lower affinity, explaining why this has the same growth pattern as maltose. All strains possess SUC2, the hydrolase responsible for inulin degradation so that does not explain the different phenotypes (Gascón, NP and JO, 1968). Though the same pattern of growth is seen on raffinose for which utilisation ability is also conferred by SUC2. The apparent carbon catabolite repression phenotype has been reported (Neigeborn and Carlson 1984), finding SNF6 to be important in sucrose non fermenting strains. It was determined that certain mutations of SNF6 also prevented utilisation of galactose and maltose by production of non-glycosylated invertase leading to a glucose-repressed phenotype. Genomic data shows that only S288c has a copy of SNF6. Alignment of the SNF6 sequences reveals that the Ethanol Red has 4 mutations within the SNF6 gene which may explain the differences. S288c, is annotated as encoding for neither GAL2 or GAL7 which are both involved in the galactose utilisation pathway (Douglas and Condie 1954). S288c lacks the gene AGA2 which is an alkaline α-galactosidase which is involved in the conversion of stachyose to raffinose which would explain its lack of growth on this carbon source. Additionally, this is linked to SUC2 as raffinose metabolism would also be required to utilise stachyose. The genomic ability for the utilisation of arabitol was investigated, it has been found that the gene ArDH1 is responsible for arabitol metabolism in *C. albicans*, however, no homologues to the gene were found in Ethanol Red by BLAST search.

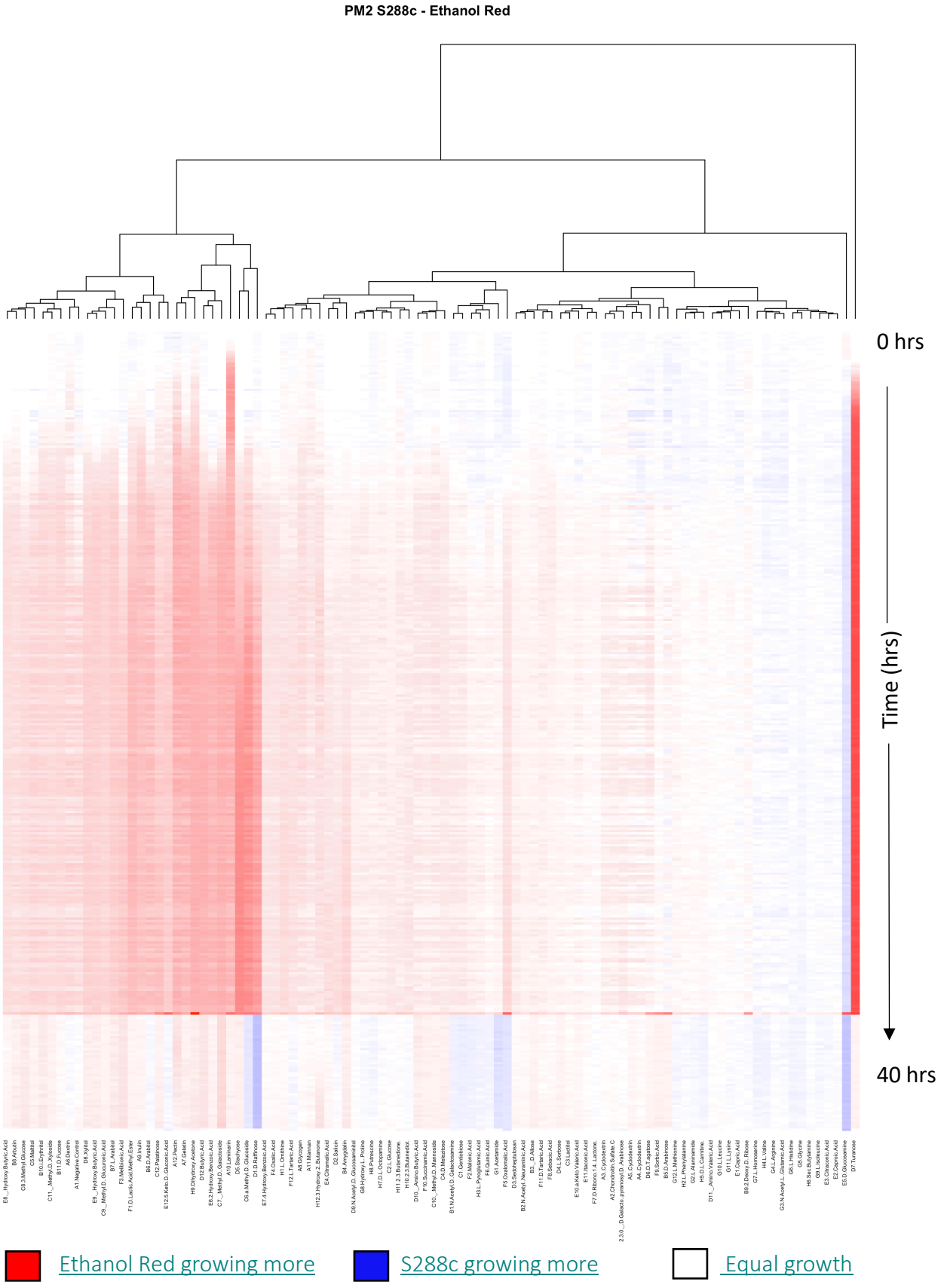


Figure 4.8. Heat map showing the growth of Ethanol Red and S288c on PM2 carbon. All parameters and colours scheme the same as figure 4.2.

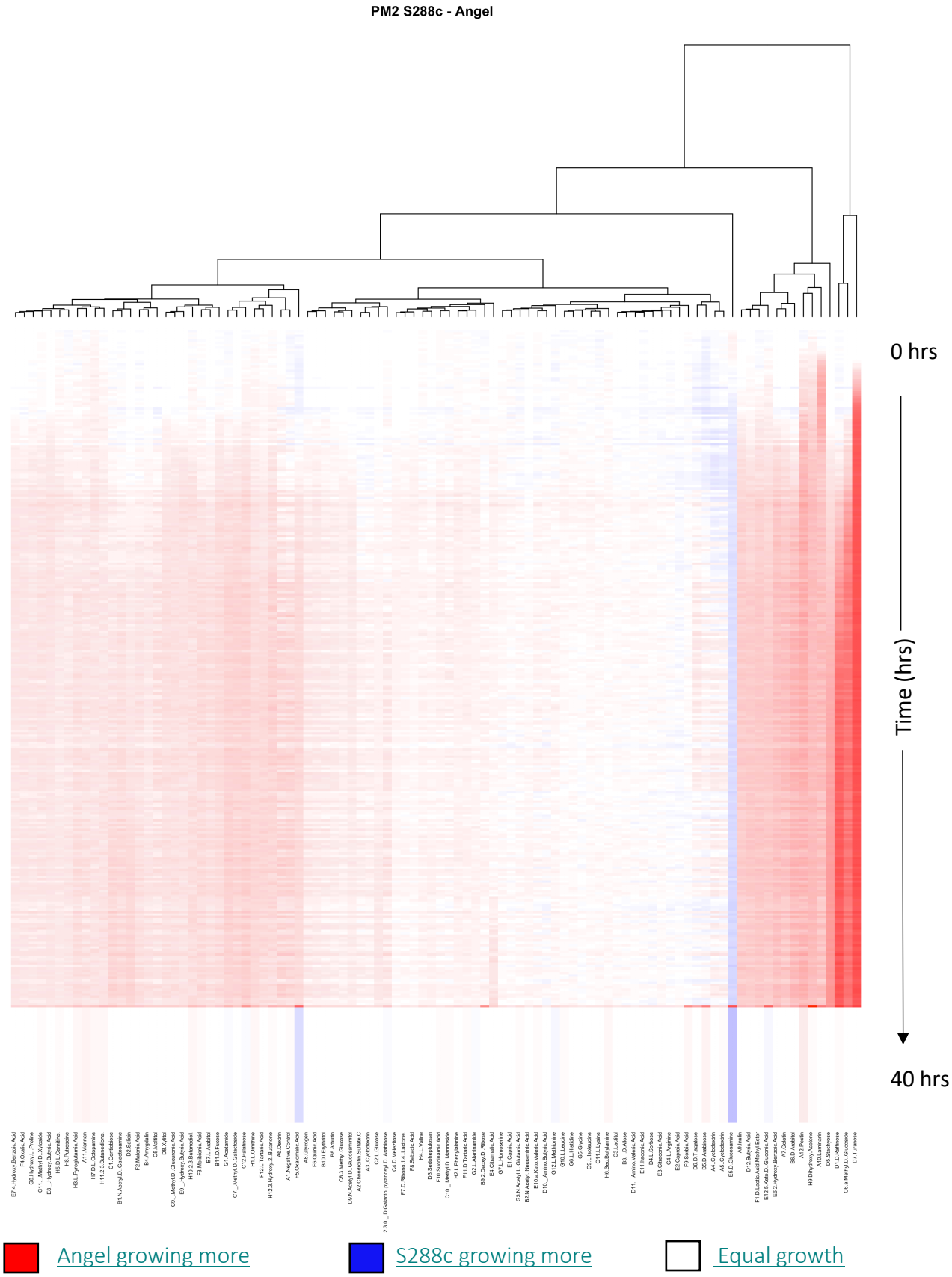


Figure 4.9. Heat map showing the growth of S288c and Angel on PM12 carbon. All parameters and colours scheme the same as figure 4.3.

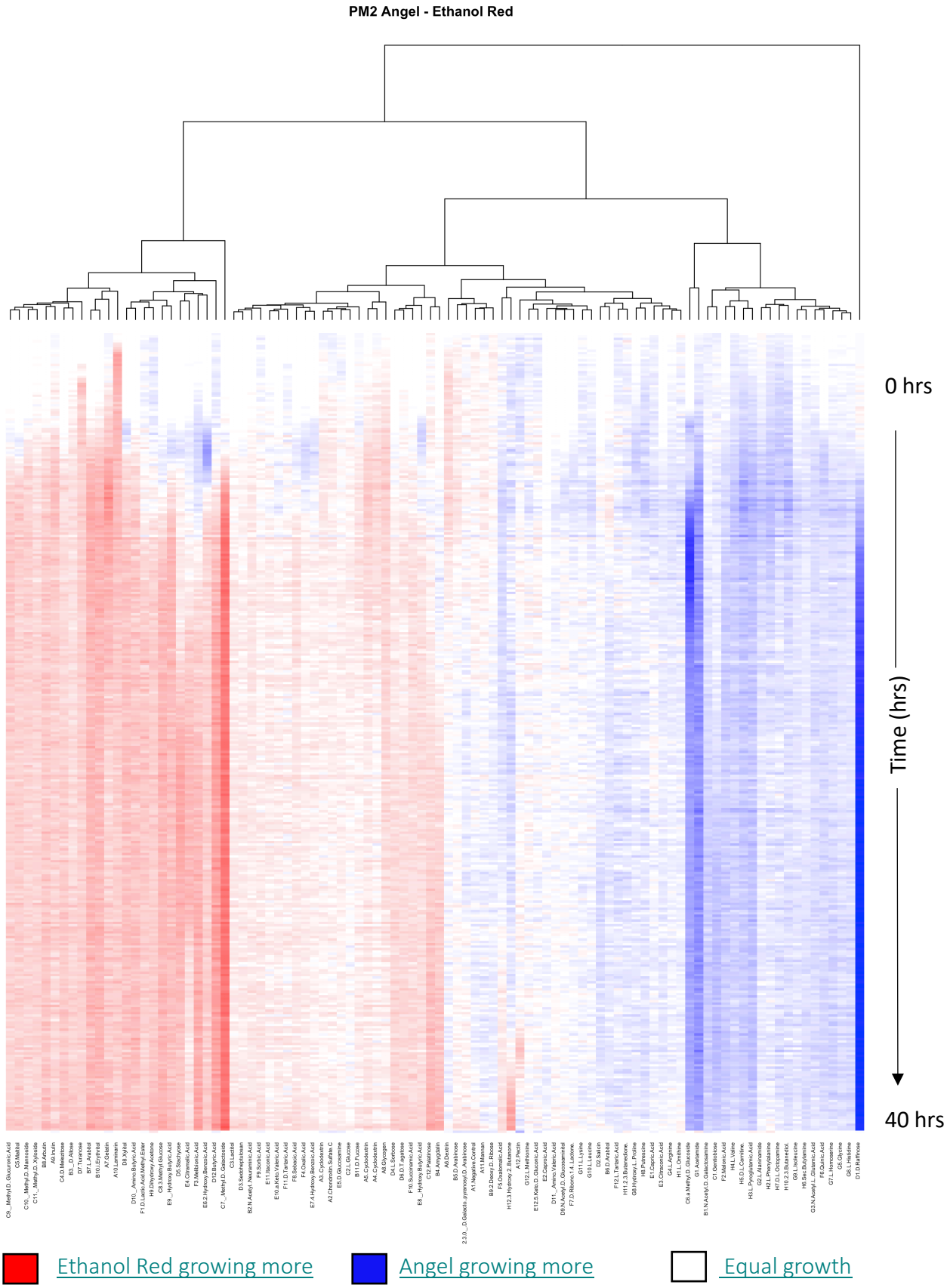


Figure 4.10. Heat map showing the growth of Ethanol Red and Angel on PM2 carbon. All parameters and colours scheme the same as figure 4.4.

4.2.2.1.4 PM5 Nutrient Supplements

In the nutrient supplement plate all conditions bar two in the case of Ethanol Red and three in the case of Angel, the industrial strains outperform S288c. This could suggest that the industrial strains have more ability to grow to a higher density or to respire more at the same density than S288c when not in the presence of any inhibitors. This is supported anecdotally where in all experiments involving growth of these three strains, S288c has reduced growth. As well as in the context of the conditions of the Biolog in which PM1 contains a well with D-Glucose as the carbon source. Since this is the preferred carbon source of *S. cerevisiae* this can be used as a positive control. Using D-Glucose as a carbon source for Angel and Ethanol Red the measured growth based on respiration is 239 and 235 respectively while for S288c it is only 186. Since normal growth on a preferred carbon source only gives S288c 78% of the growth Ethanol Red and Angel achieved, it is possible that the effect seen on the nutrient supplements is due to an inherent ability of the industrial strains to grow more. Conversely to what is seen in most of the other plates when Ethanol Red and Angel are compared to each other, the general trend is that of Angel having improved growth. In most conditions this takes the form of Angel growing faster initially only to be matched by Ethanol Red only a few hours later. In a few conditions Angel outperforms Ethanol Red until the end point, most notably biotin but also tryptophan, deoxy inosine, hematin, hypoxanthine, methionine, histidine. ARO8 and ADE5 have both been reported to affect a strains ability to utilise methionine, genome sequences show that S288c does not contain these genes while the industrial strains do (Jones and Fink, 1982).

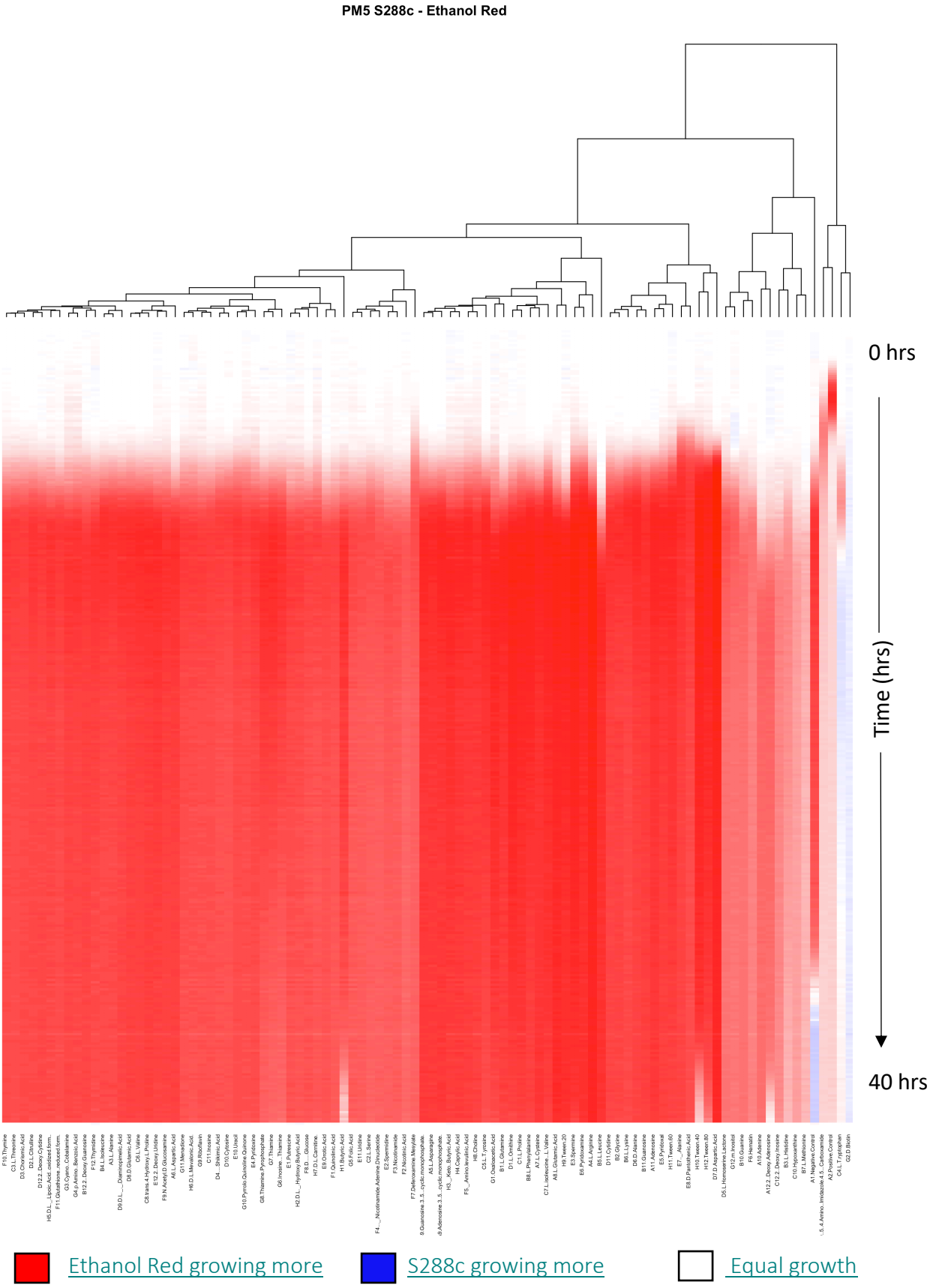


Figure 4.11. Heat map showing the growth of Ethanol Red and S288c on PM5 carbon. All parameters and colours scheme the same as figure 4.2.

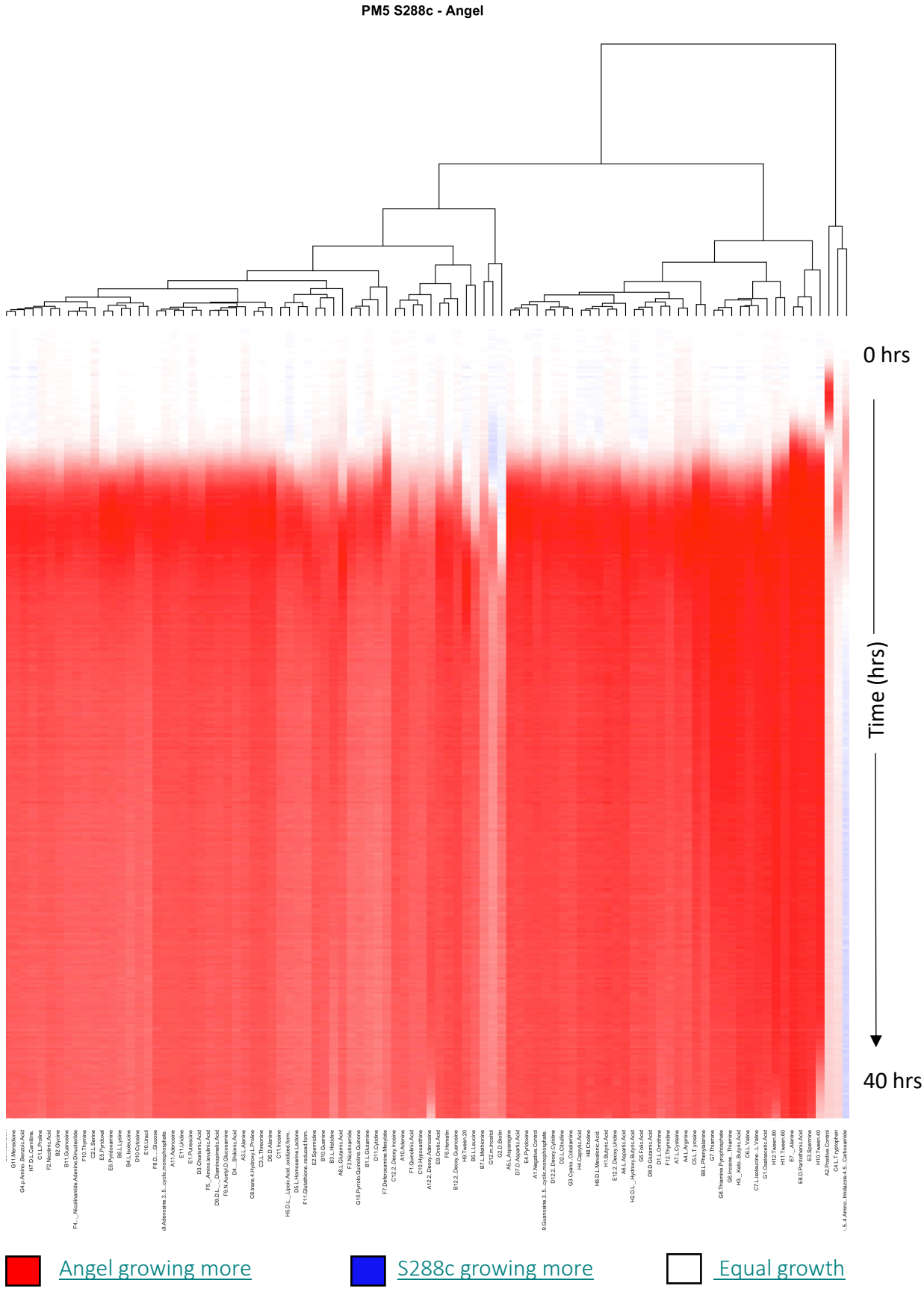


Figure 4.12. Heat map showing the growth of Angel and S288c on PM5 carbon. All parameters and colours scheme the same as figure 4.3.

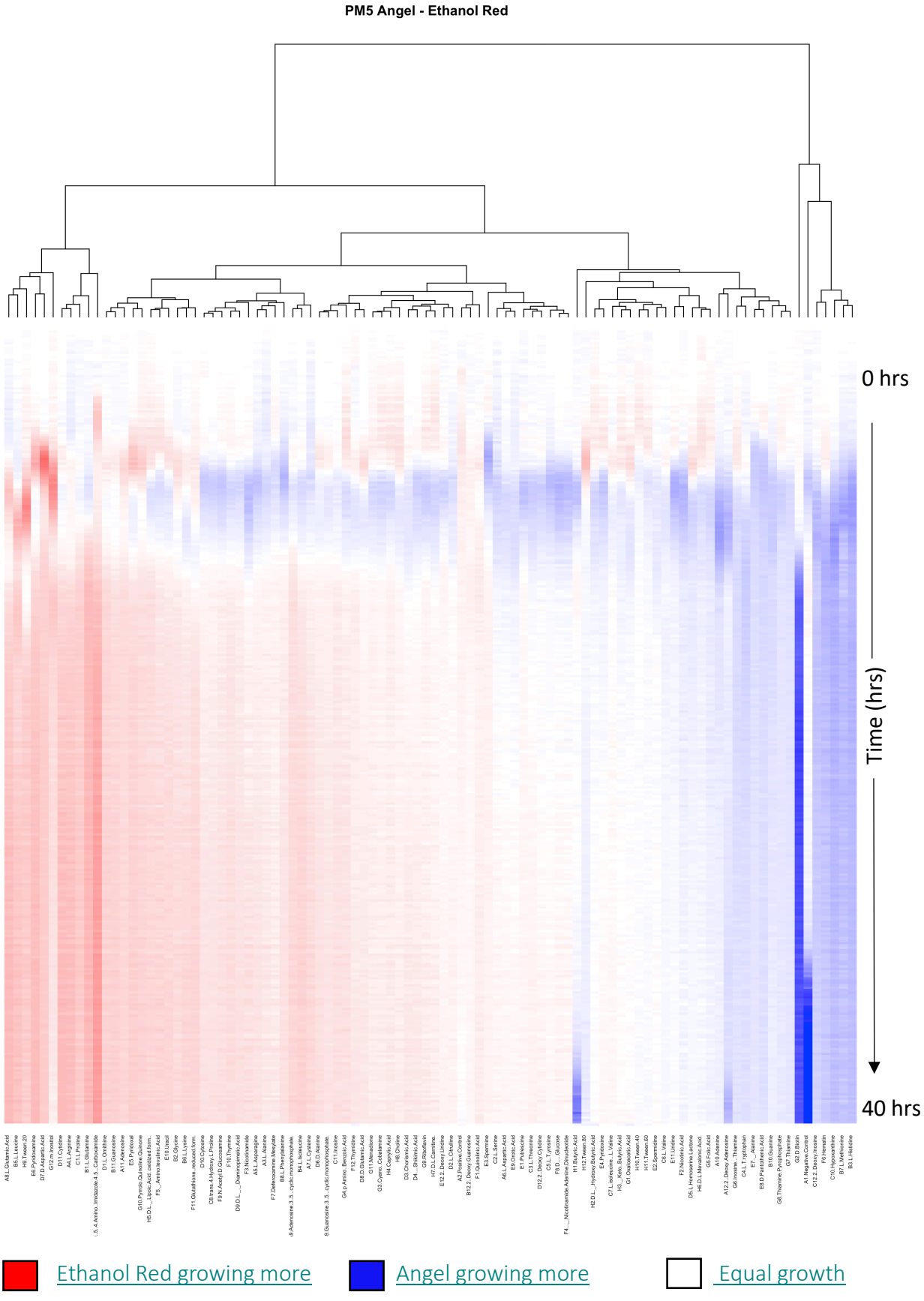


Figure 4.13. Heat map showing the growth of Ethanol Red and Angel on PM5 carbon. All parameters and colours scheme the same as figure 4.4.

4.2.2.1.5 PM9 osmolytes

Similar to the nutrient sources plate, the results observed from the growth on various osmolytes suggest that many of the differences could be explained by the higher cell density achieved by the industrial strains over this time in normal conditions. For this set of conditions, the industrial strain completes its exponential growth phase faster than S288c, but the lab strain does catch up before the end of the growth period. There are a number of cases here where S288c outperforms the industrial strains. S288c outperforms both industrial strains on sodium phosphate. It also outperforms Ethanol Red on urea and ethylene glycol.

Ethanol Red has slightly improved growth on a number of conditions but particularly on sodium chloride and sodium lactate. Angel also has slightly higher levels of growth in several conditions when compared to Ethanol Red but none that are particularly pronounced. When the two industrial strains are compared to each other it can be seen that in most conditions both grow to a very similar level though overall Ethanol Red slightly outperforms Angel. On sodium phosphate, sodium lactate and sodium chloride, Ethanol Red has markedly more growth than Angel.

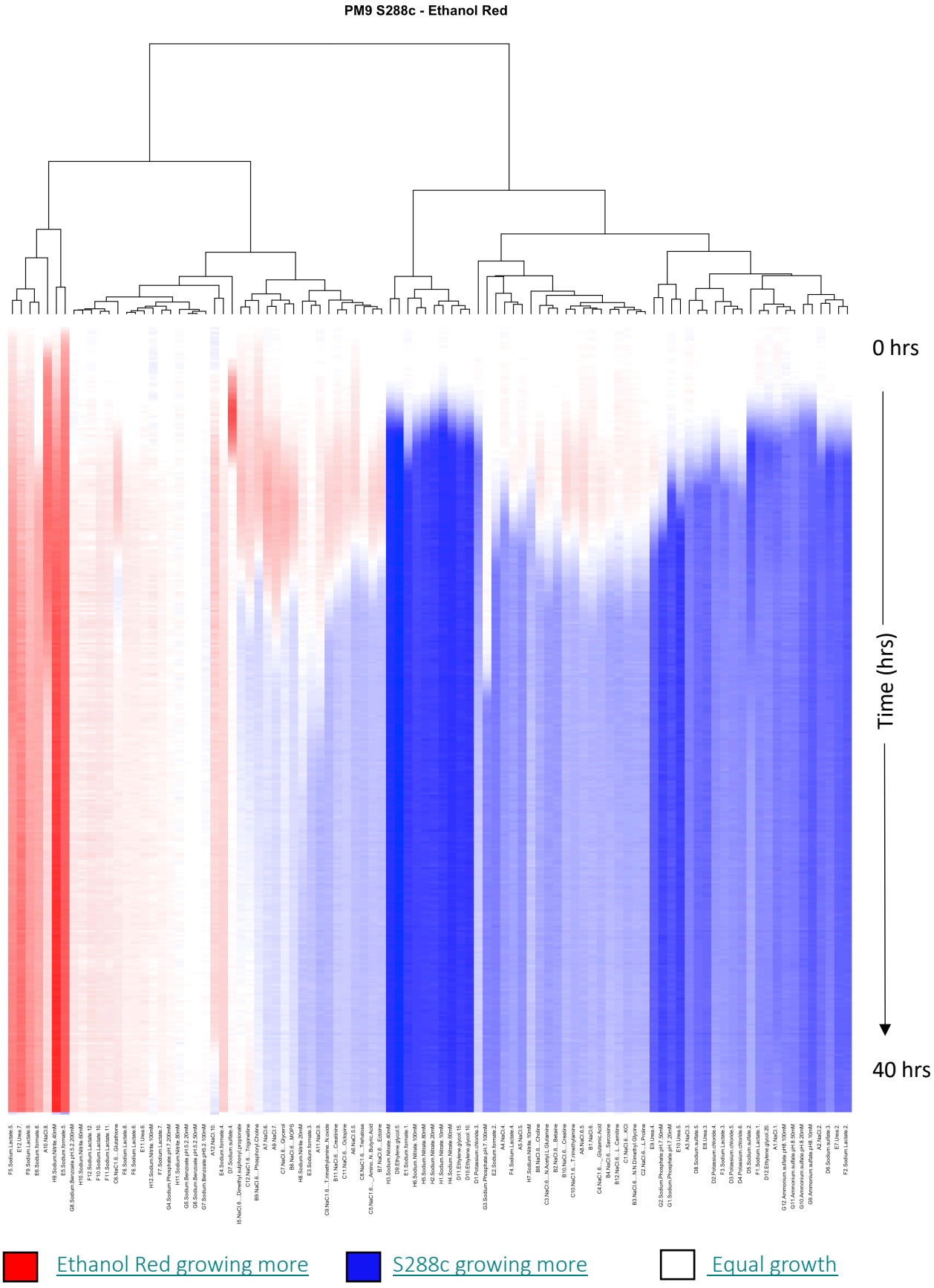


Figure 4.14. Heat map showing the growth of Ethanol Red and S288c on PM9 osmolytes. All parameters and colours scheme the same as figure 4.2.

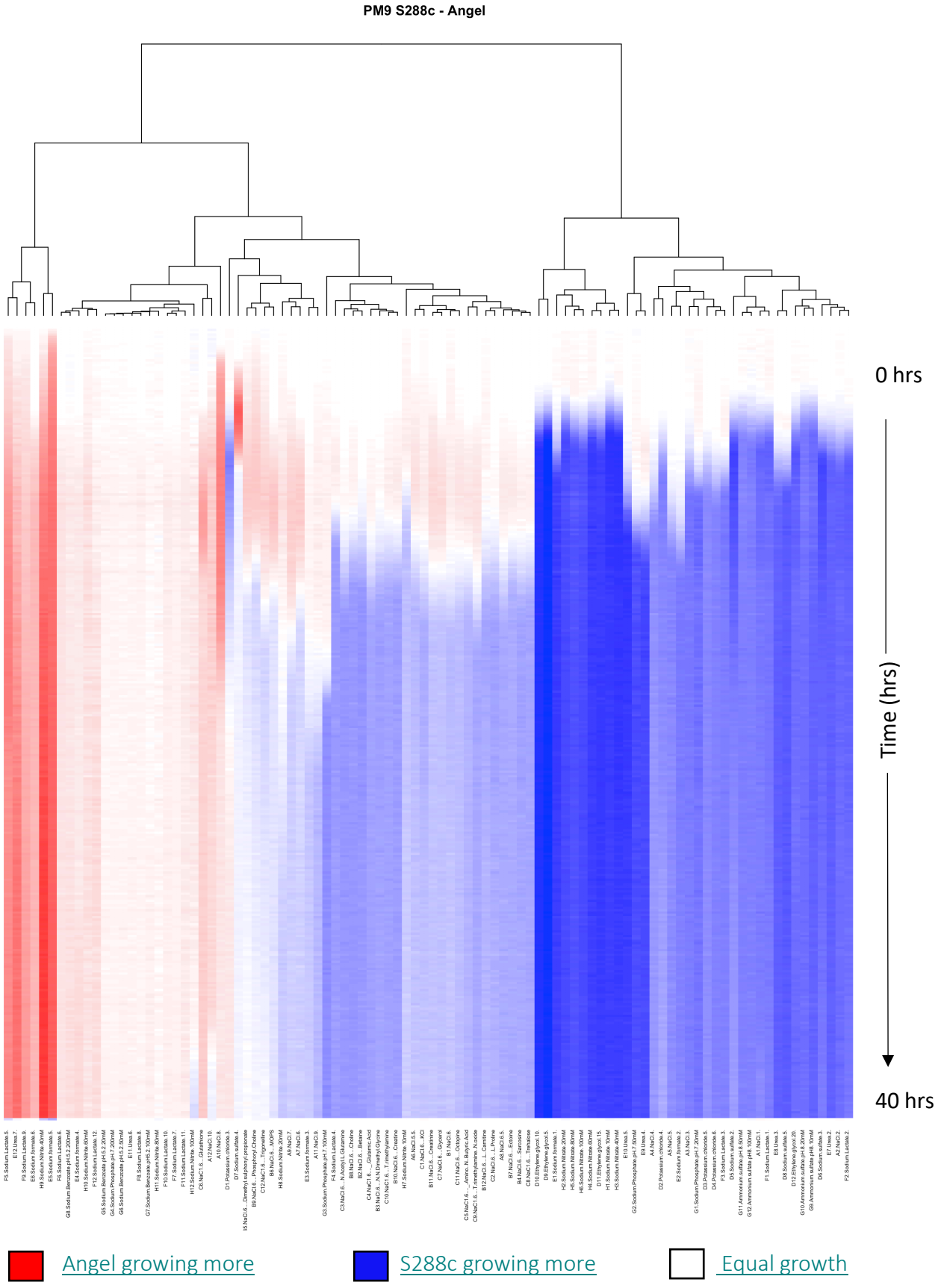


Figure 4.15. Heat map showing the growth of S288c and Angel on PM9 osmolytes. All parameters and colours scheme the same as figure 4.3.

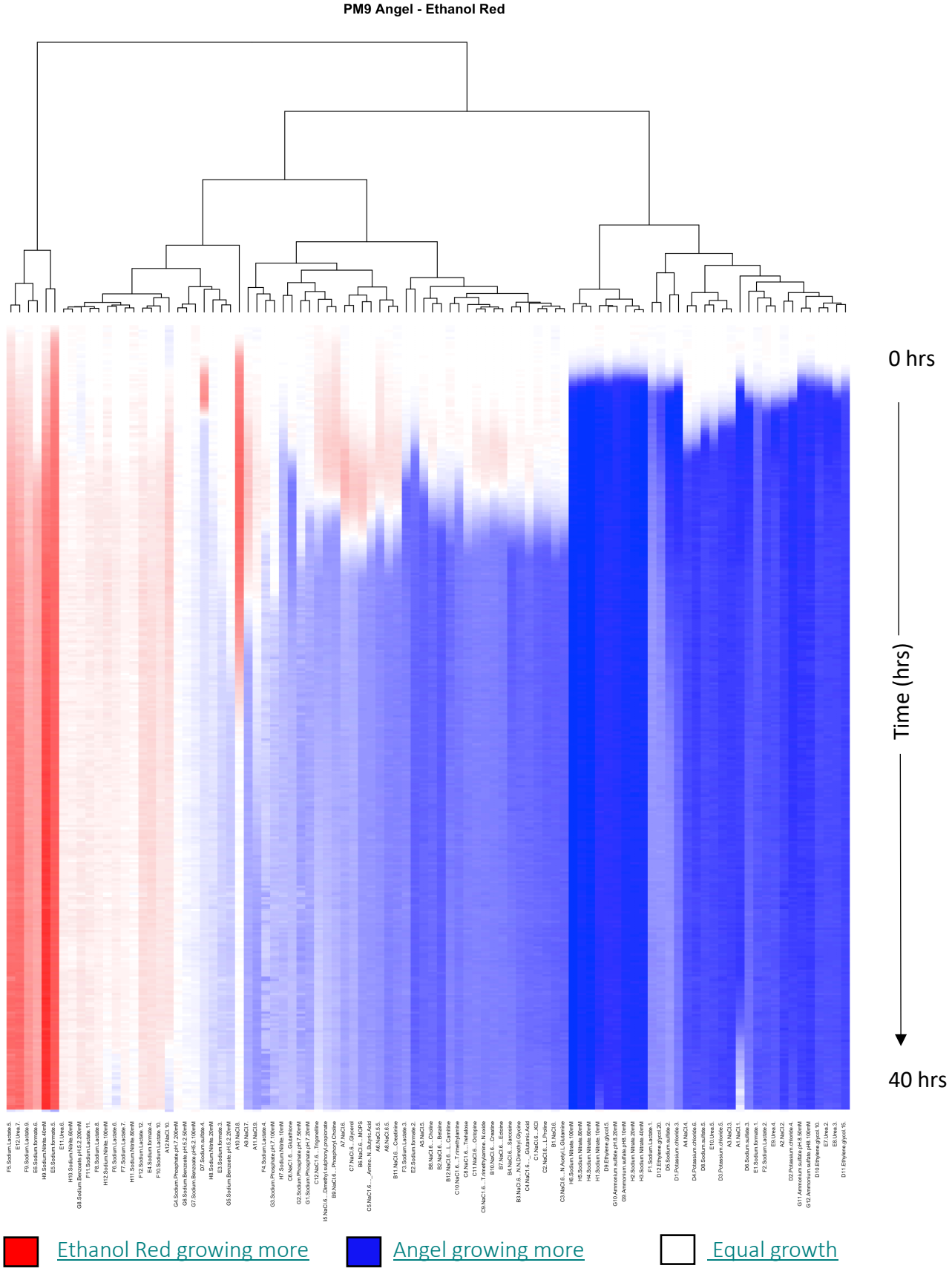


Figure 4.16. Heat map showing the growth of Ethanol Red and Angel on PM9 osmolytes. All parameters and colours scheme the same as figure 4.4.

4.2.2.1.6.PM10 – pH

PM10 provides several different pH levels for growth, this also includes a number of amines that may act to protect against an adverse pH. As with the other plates utilising rich media there is a growth lag with S288c, which can be seen with the purple colour being visible near the beginning of many growth curves and before S288c catches up and the middle of the spectrum pink colours are displayed again. There are a few cases where the Ethanol Red continues to outperform, these are pH 4.5 L-asparagine, and pH 9.5 L-glutamic acid. Angel also grows more than S288c on pH 4.5 L-asparagine as well as on pH 9.5 L-proline and pH 9.5 L-norvaline. This is suggestive that the ability to grow on both acid and alkaline pH can be facilitated by certain amines to a greater extent in the industrial strains compared to S288c. When the two industrial strains are compared Angel seems to outperform Ethanol Red in several conditions including pH 9.5 L-threonine, pH 9.5 L-norleucine pH 5, pH 3.5, pH 9.5 L-serine and pH 4. There are only two conditions in which Ethanol Red outperforms Angel on this plate and those are pH 9.5 L-methionine and pH 6.5.

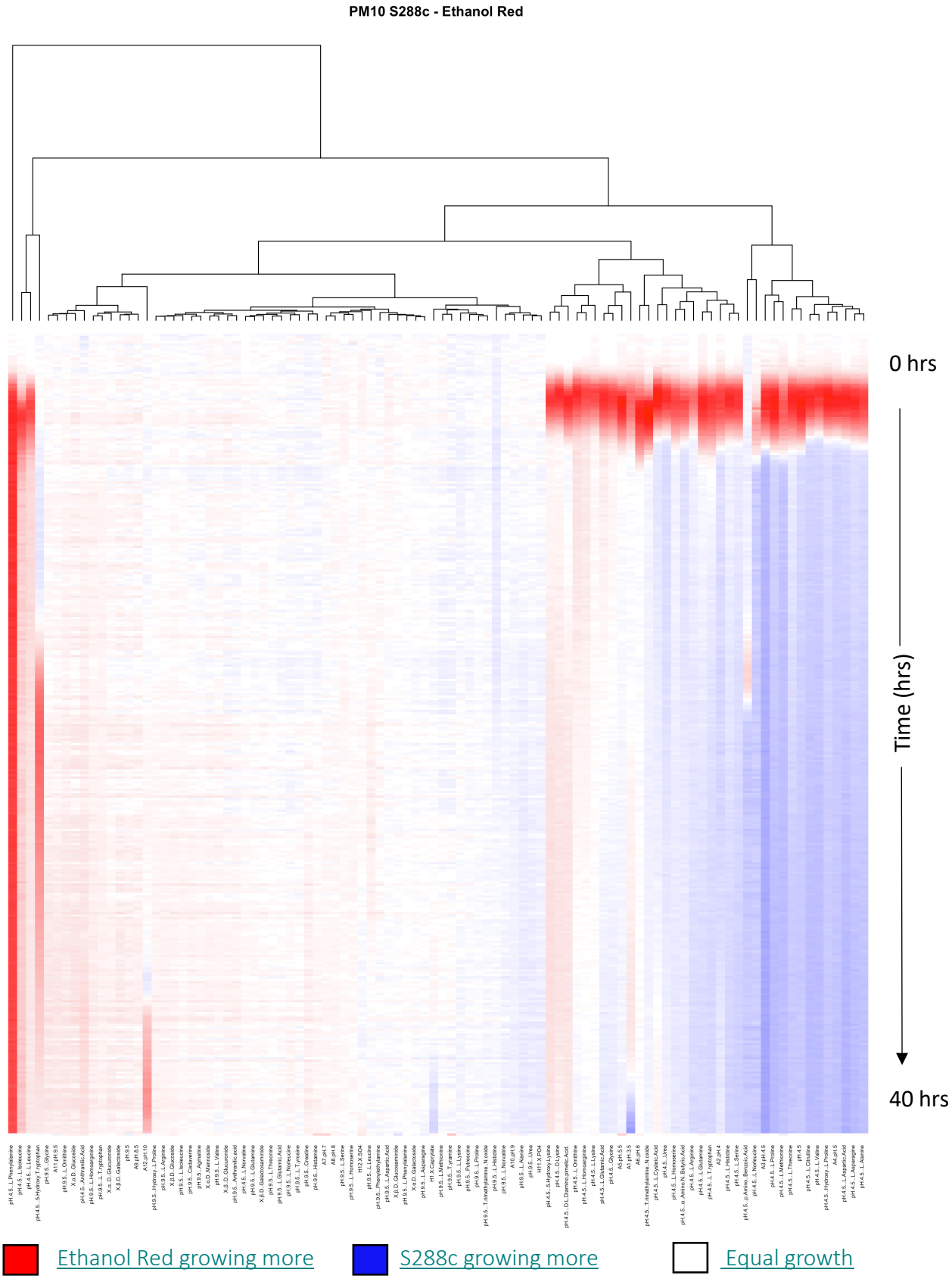


Figure 4.17. Heat map showing the growth of Ethanol Red and S288c on PM10 pH. All parameters and colours scheme the same as figure 4.2.

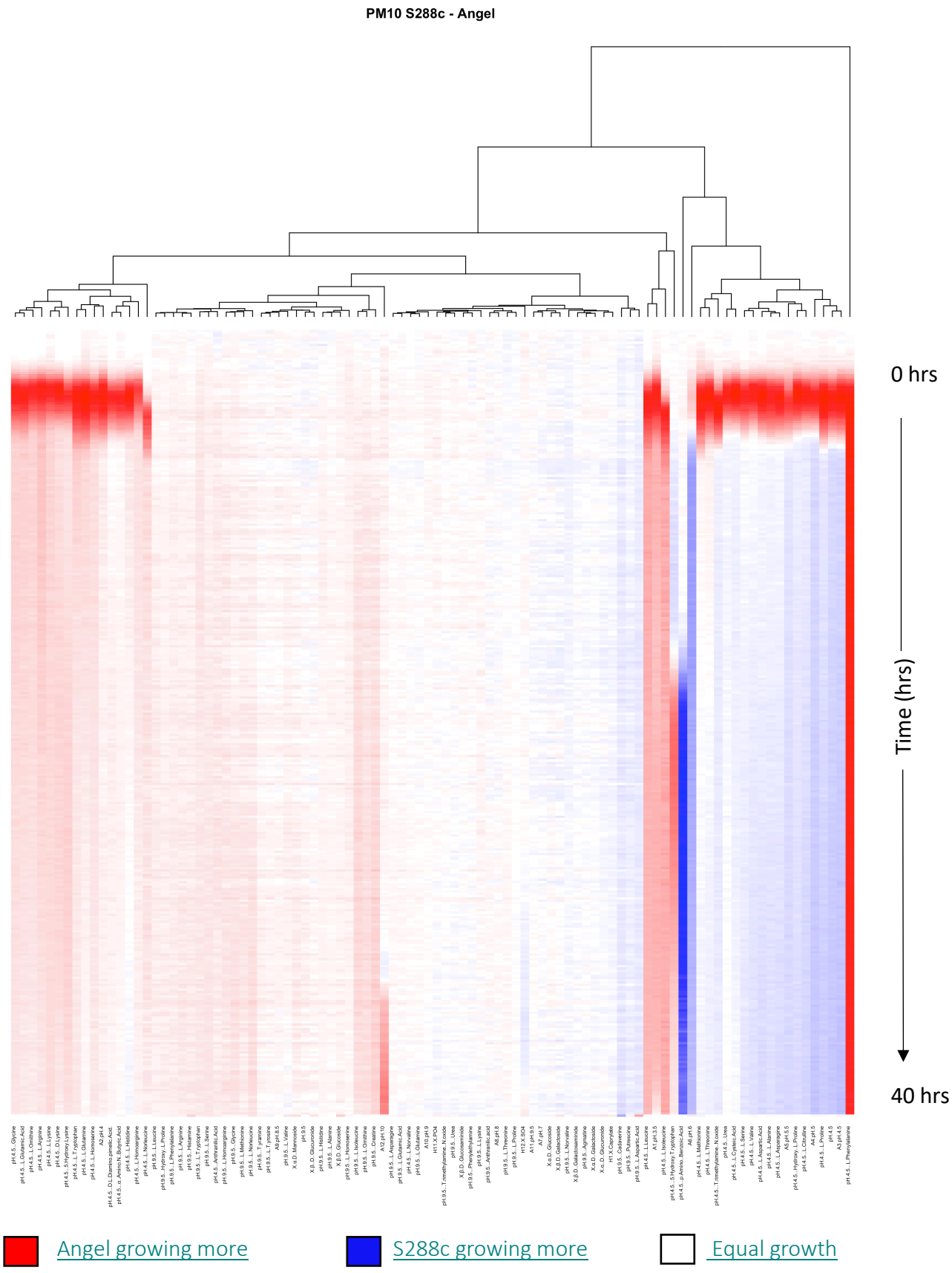


Figure 4.18. Heat map showing the growth of S288c and Angel on PM10 pH. All parameters and colours scheme the same as figure 4.3.

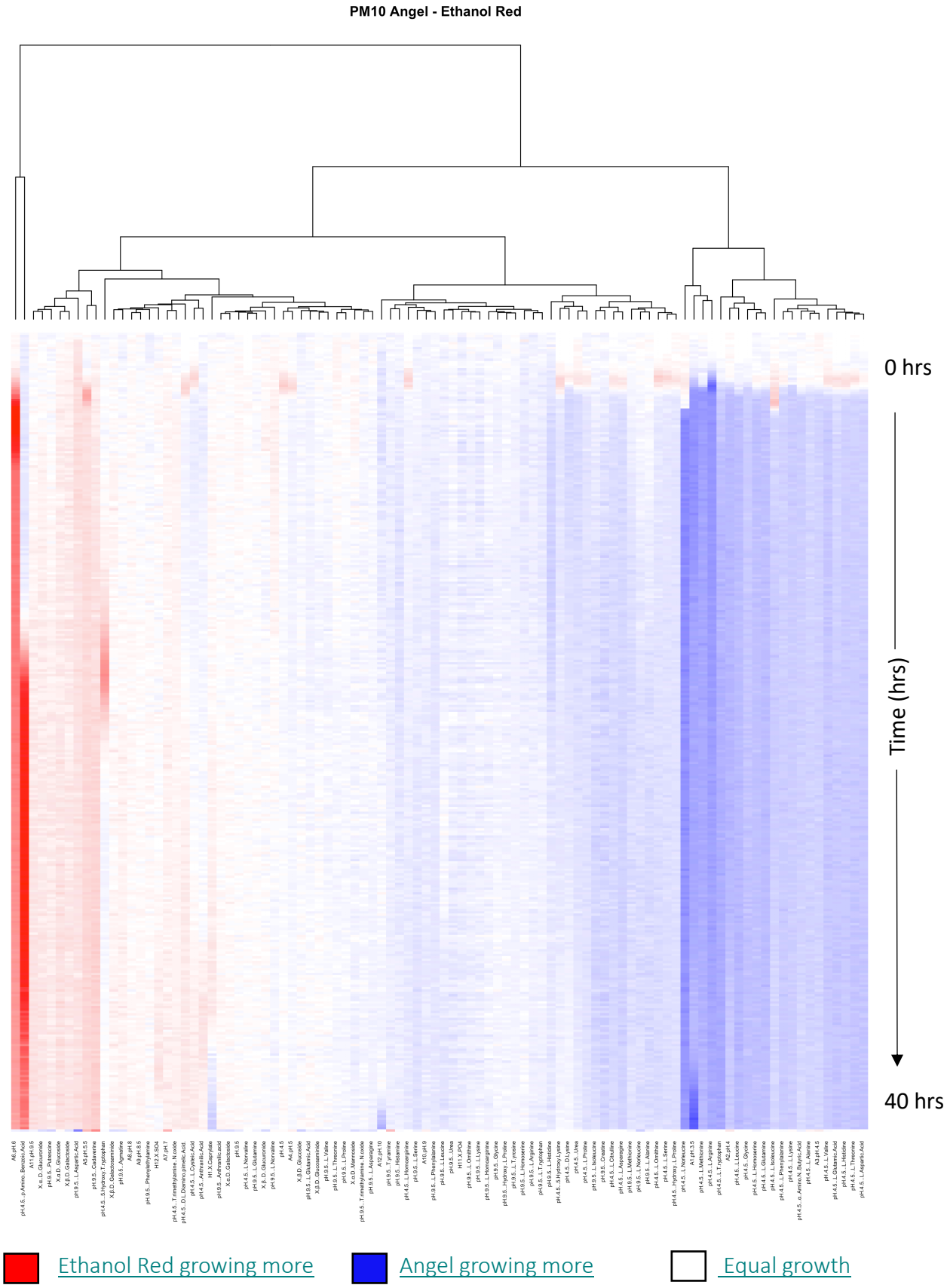


Figure 4.19. Heat map showing the growth of Ethanol Red and Angel on PM10 pH. All parameters and colours scheme the same as figure 4.4.

4.2.2.3 Genomics

The strains discussed phenotypically here were also sequenced as part of this work or, in the case of S288c, already publicly available in complete form. This allowed us to analyse the phenotypic data with the genomic information as context. The availability of both sets of information results in more relevant conclusions being drawn, as has been noted the presence or absence of a gene implicated in a particular phenotype does not always result in the phenotype when tested.

The results from the SNP calling for the strains assessed phenotypically will also be discussed here in order to determine the correlation between phenotype and genotype seen. In the case of the two industrial strains compared to S288c there were too many SNPs to be resolved in a graphical format for visualisation. As such sequencing data from chapter three was used to complement the analysis of these data

4.3 Conclusions

This work presents a simple and cost-effective workflow for high quality sequencing that can be performed in-house by industry allowing for rapid result turnaround and intellectual property protection. This data also shows where this method is effective, high coverage for SNP calling and where It falls short, low telomere coverage in scaffolded strains. Despite this, the assembly of Ethanol Red provided here is a marked improvement on the best currently available public sequence. It has also been shown here that a complete genome can greatly improve the assembly of other similar strains meaning that the cost and time involved in sequencing can be reduced.

High throughput phenotypic data can be utilised as a method for rapid screening of candidate strains in hundreds of conditions simultaneously. This data shows that there is less variation than expected between the industrial strains and the lab strains in terms of carbon utilisation. With many of the more widespread phenotypic differences being seen on other plate types. Also observed here is that the genotype and phenotype don't always correlate directly as might be anticipated. For this reason, it is thought that both types of data are valuable, but more so together. With both of these data sets, together, a fuller understanding is gained that in turn leads to improved future experimental work.

Some limitations of this type of phenotypic data are that each culture is assessed individually, which is vital for understanding of pathways but further work to understand how each strain behaves in a mixed culture would provide additional useful information. For example, in the case of carbon catabolite repression, genomic differences in the regulators may lead to one strain utilising a mixed carbon source more fully than another.

In the context of industry where the resulting phenotype is important above all else the genomic information can provide vital answers in order to make targeted phenotypic changes in a lineage. This strategy, whilst optimised here for a Eukaryote, can be used to provide information on all types of organisms. As is shown here, the techniques are now robust enough to easily sequence relatively complex organisms and so the scope is wide for use with a variety of industrially relevant microorganisms.

Chapter 5. Whole-genome Sequencing of an Industrial Lineage of *Streptomyces*

5.1 Introduction

5.1.1 *Streptomyces* producing clinically relevant compounds

As discussed in the main body of the introduction *Streptomyces* are known to produce a vast number of clinically relevant compounds that are of interest to industry both with a view to use *Streptomyces* strains for drug discovery due to their novel chemistry and as production organisms. One of the aspects of S*treptomyces* biology that has become clearer in the genomics era is the complex phylogeny in the actinomycete genera. As more genomic information becomes available it is increasingly common for strains or species to be reclassified. In addition to this, new culturing methods and metagenomics adds variety to the pool of whole-genome sequences of organisms that we have and often allows for definition of new and novel species. As such when studying *Streptomyces* or indeed any microorganisms at this level, historical classification which has been based on phenotype often only gives limited insight to their true placement and as such should be viewed with scepticism. Furthermore, due to the prevalence of horizontal gene transfer amongst this genus, particularly with biosynthetic gene clusters, the phylogeny often doesn't give a particularly clear indicator of a strain's biosynthetic potential. This coupled with the difficulty of inducing biosynthetic gene cluster expression under laboratory conditions means that whole-genome sequencing is currently the best method to determine the biosynthetic potential of a strain. As such this chapter will focus on the sequencing and analysis of four *Streptomyces* strains that are of industrial interest and that have limited genomic data available. They are *Streptomyces ardus* (DSM 40603) (MED001), *Streptomyces lavendulae* (*S. caespitosus*) NRRL 2564 (MED002), *Streptomyces caespitosus* AN-9 (NRRL 12513) (MED003) and *Streptomyces caespitosus* T-17-135 (NRRL 12508) (MED004).

5.1.1.1 Production of the Mitomcyin Group of Compounds

Of the strains analysed here, *S. lavendulae* is the most studied and the only one with a published genome at the time of writing (August 2020). The strain *Streptomyces lavendulae* was first isolated from soil in 1916 and was initially of interest because it was found to produce streptothricin (Waksman and Woodruff, 1942). The discovery of Streptothricin from *S. lavendualae* was made in 1942 and is historically important because it was the first basic and water soluble antibiotic to be isolated (Waksman, Harris, and Lechevalier 1951). Since then several other antibiotic compounds related to streptothricin have been found though their use in the clinic has been hampered due to toxicity in humans (Khokhlov, 1983). Another molecule that is produced by this same strain is of more interest for its use as an anti-cancer compound. Mitomycin C is a chemotherapeutic drug used to treat numerous cancers including papillary bladder tumours (Serretta *et al.*, 2016). Mitomycin C is also produced by several other *Streptomyces* including *S. ardus* and *S. caespitosus*. The mitomycins were first discovered by Hata et al. in 1956 from *S. caespitosus*(Hata et al. 1956; Hata and Sugawara 1956) and Porfiromycin was first isolated from *S. ardus* in 1960 by DeBoer et al. (DeBoer, C., Dietz, A. and Lummis, 1960). The mitomycins all have a core structure in common containing an aziridine ring, pyrole (1,2-a) indole ring and a pyrrolizidine reside. They also contain three carcinostatic groups which are aziridine, quinone and urethane (Kersten, 1975). The structures of mitomycin A, B, C and porfiromycin is given in figure 6.1.

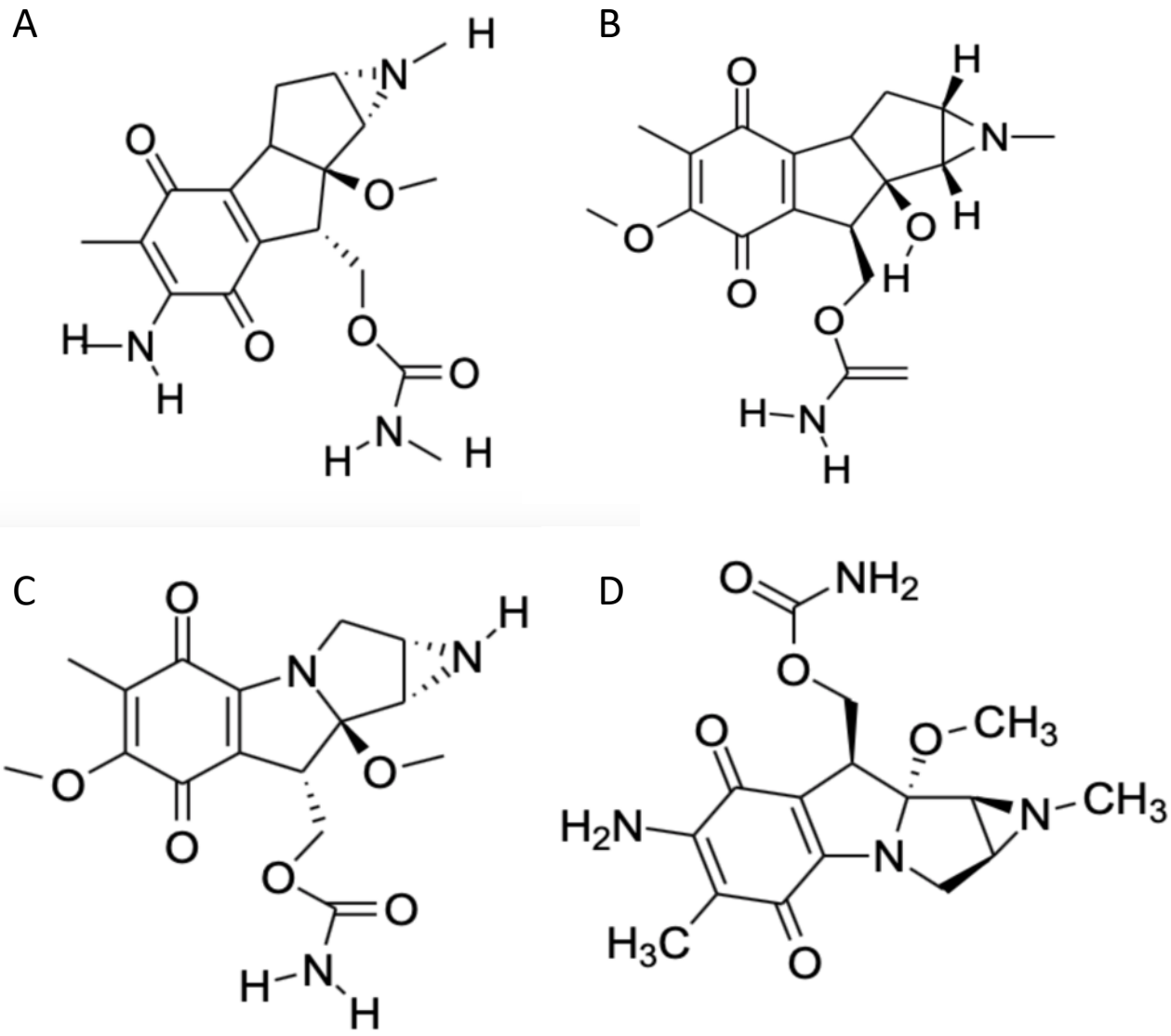


Figure 5.1. The chemical structure of the mitomycin family compounds. A) Mitomycin A. B) Mitomycin B C) Mitomycin C and D) Porfiromycin

5.1.1.2 Mitomycin Mechanism of Action

The mitomycins work by bio-reductive alkylation, which is an uncommon mechanism to these drugs that describes the requirement for them to be reduced *in vivo* before being active and binding to DNA by alkylation which results in the formation of cross links in the DNA double helix (Vijayaraj Reddy and Randerath, 1987). Mitomcyin C is the most well studied and is active in hypoxic environments meaning it is useful in various types of cancer including breast, non-small cell lung, pancreatic and bladder cancers. However,it also has numerous potential side effects which limit it from having wider clinical applications (Bass *et al.*, 2013). The two main side effects are leukocytopenia and thrombocytopenia which are decreased white blood cells and platlets respectively (Verweij and Pinedo, 1990) with multiple other less common but still relevant side effects.

5.1.1.3 Mitomycin Biosynthesis

The biosynthesis of mitomycin C is relatively well understood. It is controlled by cluster of 47 genes covering a total of 55 kilobases in *S. lavendulae* shown in figure 6.2 (Ji *et al.*, 2009). The gene cluster was elucidated and confirmed in 1999 where a series of deletions were performed to decipher which genes were required for the production of mitomycin (Yingqing, Varoglu and Sherman, 1999). Despite this there remained many genes of unknown function. Since the formation of mitomycin is dependent upon (3-amino-5-hydroxybenzoic acid AHBA), there are 5 genes involved in AHBA biosynthesis throughout this cluster, as well as numerous genes that are responsible for side-group modification which contribute to the formation of the final product. There are also four genes involved in or suspected to be involved in resistance to mitomycin. These are *mitR, mmcM, mrd* and *mrc*. Mrd binds to mitomycin at a ratio of 1:1 and Mct which is encoded 5 kbp upstream of the cluster transports the bound mitomycin out of the cell. It was found that co-expression of both of these genes, but not either individually in *E. coli* significantly improved resistance to mitomycin (Sheldon *et al.*, 1999).

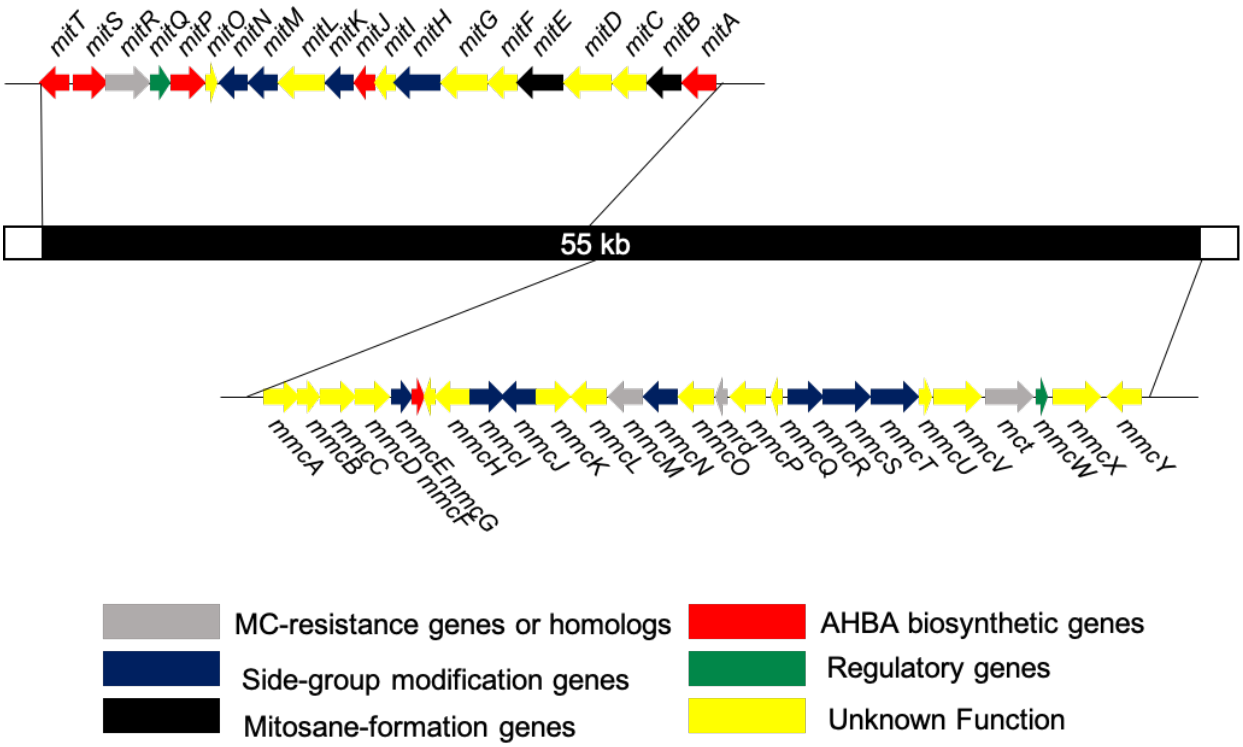


Figure 5.2 Mitomycin gene cluster. Showing the gene cluster involved in mitomycin C biosynthesis from *Streptomyces lavendulae* adapted from Sheldon et al. 1999

5.1.2 Aims and Objectives

The differential expression of mitomycin C observed between these strains of similar phylogeny suggests that there is potential for molecular biology-based strain improvement. Prior to this work whole-genome sequencing of Streptomyces was more problematic than for other organisms such as yeast, *E. coli* and other bacteria owing to their extremely high GC content and so it was desirable to develop a method for effective sequencing of high GC content organisms in-house. Whole-genome sequencing in this case will provide a fuller depth of information on the strain and allow us to analyse the biosynthetic potential of each strain. It is hypothesised that high quality genome assemblies will allow for a targeted molecular biology strain improvement protocol to be generated that will improve the production organism for this compound of interest.

5.2 Results

5.2.1 Sequencing output and quality assessment

Four strains that are potential Mitomycin C producers were investigated for genomic differences that may explain their varied phenotypes in terms of their ability to produce the compound of interest.MED001 and MED002 both produce Mitomycin C at different levels whilst MED003 and MED004 produce high levels of mitomycin A, a precursor in the pathway to mitomycin C but no mitomycin C is detected.

The Illumina Mini-Seq was used to perform whole genome sequencing and the four strains were analysed bioinformatically. Illumina paired end, 150 bp read, sequencing was used with an average fragment size of 600 bps. After sequencing on the Illumina MiniSeq, the quality of the reads was assessed by FastQC (Andrews, 2014) and the quality was high throughout and trimming of the reads before assembly was not carried out. Quality of raw reads is shown in figure 6.3.

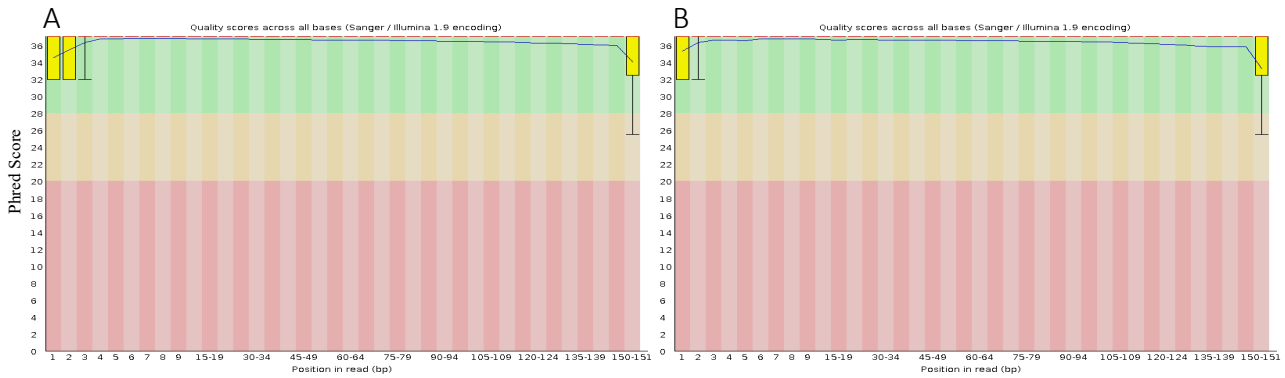


Figure 5.3. FastQC output for a representative set of Illumina Mini-Seq reads for the strains of *Streptomyces lavendulae.* A. MED001 forward reads. B. MED001 reverse reads. Each graph shows the position along the read length 1-151 base pairs on the x-axis and the Phred score on the y axis. The collective quality score for all reads in each set are displayed as box plots.

5.2.2 *De novo* assembly

Once quality of the reads was determined no trimming was deemed necessary and so SPAdes was used to assemble the reads. QUAST was utilised to provide assembly statistics as a first indicator of assembly quality (Table 5.1) and the assemblies visualised using Bandage (Figure 6.4). The QUAST results indicate that the SPAdes assembles are >70 contigs of a size 1000 bp or higher for each of the four strains together with the Bandage visualisations in Figure 5.4 shows that the assemblies are fragmented. Whilst these parameters do not ensure a good or bad assembly, they are reasonably good predictors at an early stage. Using QUAST the assembly was compared to the assembly for reference genome for *S. lavendulae* (figure 5.5). It can be seen here that the genome size is considerably smaller than would be expected for this species. It can also be observed from this data that there is a considerable number of small contigs rather than most of each genome being represented in a few larger contigs as is sometimes the case. This is evidenced by the low N50 value given in table 5.1.

Parameter	MED001	MED002	MED003	MED004
# contigs (>= 0 bp)	300	291	324	255
# contigs (>= 1000 bp)	92	77	67	76
Total length (>= 1000 bp)	7066320	7065681	6291405	6305597
GC (%)	71.84	71.84	71.84	71.85
N50	177129	213739	196634	169138

Table 5.1. Selected aspects of the QUAST report on all four strains for comparison

A B

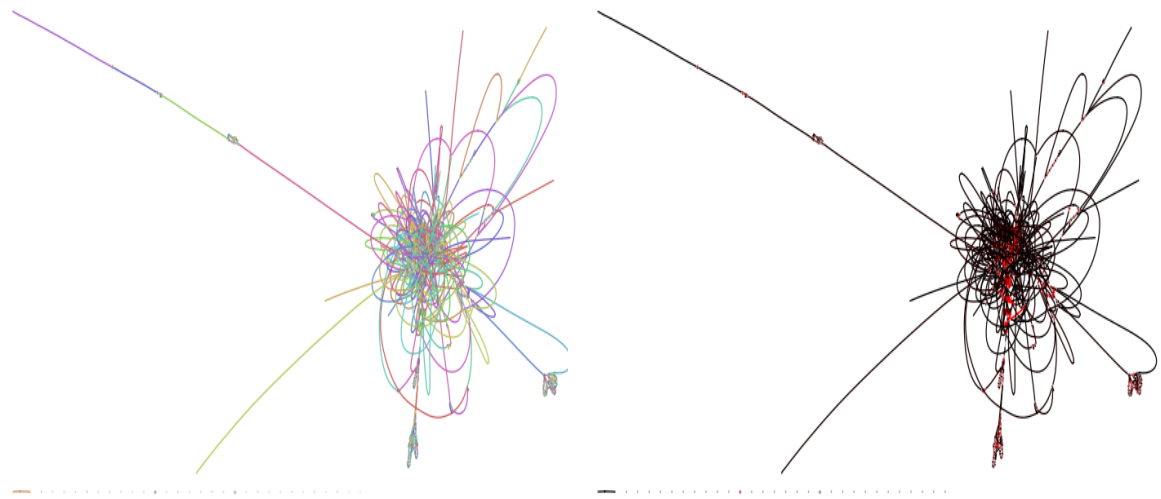
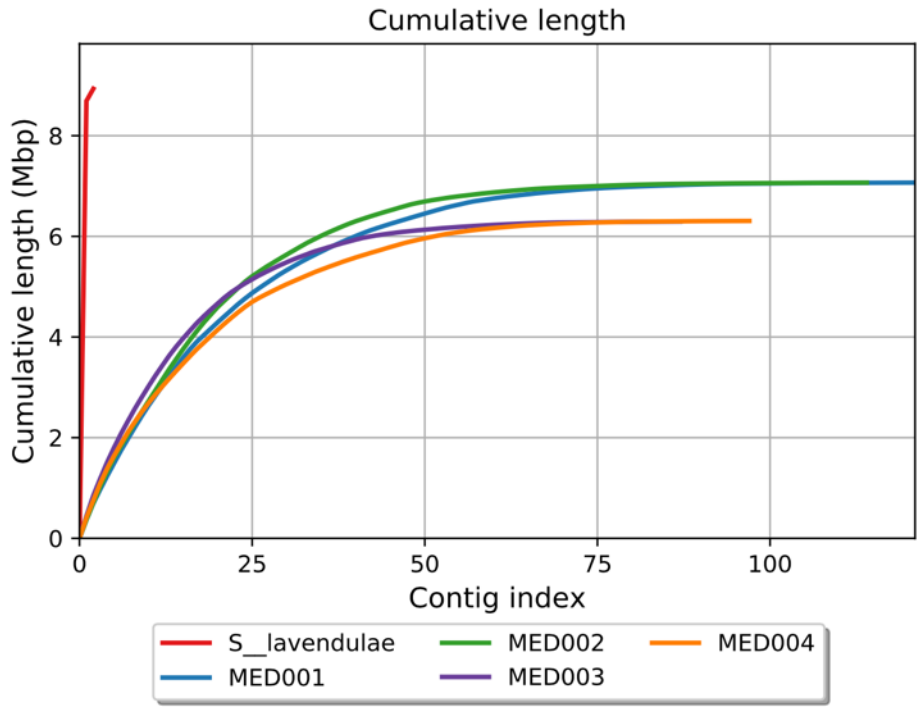


Figure 5.4. The Bandage visualizations of the assembly Graph generated by SPAdes with careful correction. A. shows the graph with random coloration on each node. B. shows the graph with colour corresponding to depth of coverage for each node.The black nodes have a depth of coverage of ~70X which corresponds to the expected coverage for the number of reads in a genome of this size. The red nodes have a coverage ranging from ~100X to 100,000X.

A



B

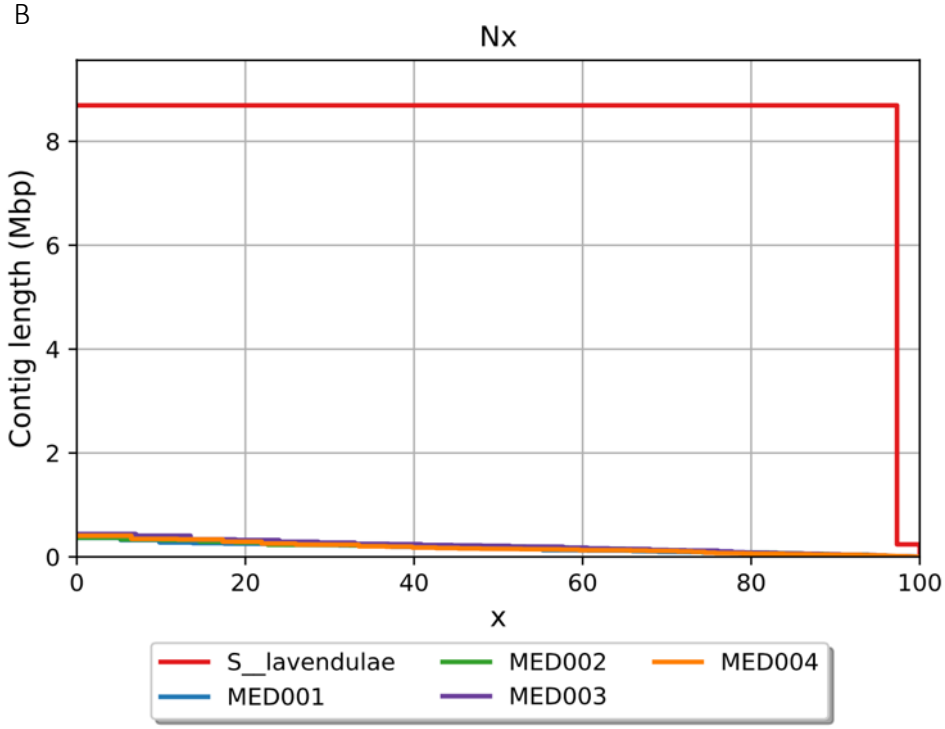


Figure 5.5. Assembly statistics graphs. Showing the assembly quality statistics as determined by QUAST of all four genomes and the reference *S. lavendulae*. A) Displays the cumulative length of each assembly with the contig number on the X-axis. Only contigs of a size of 500 bp or larger were used here. B) gives the length of each contig in the assembly

5.2.3 Reference-guided assembly

One way to improve an assembly, overall but particularly in terms of contiguity and improving the N50 value, is to use a completed genome as a reference to guide the assembly. The completed or draft reference genome acts as a scaffold that the contigs from the *de novo* assembly can be contextualised with. This can lead to a genome being in scaffolds rather than contigs in which a series of N's can be inserted between contigs to signify that they are expected to be close together based on the reference genome. This can allow for better and more informative assemblies. As there is no published genome sequence for *S. caespitosus* or *S. ardus* but there is a complete genome for *Streptomyces lavendulae* this was the obvious first choice for assembly improvement. The genome assembly used as a reference here was *Streptomyces lavendulae subsp. lavendulae* (ASM280384v1) that was sequenced by Illumina at Bielefeld University and assembled using Newbler v. 2.8 (Bekeova *et al.*, 2016). MeDuSa was used to scaffold the assembled genome using *Streptomyces lavendulae* as a reference and as can be seen in Table 5.2 this gives us more information about the ordering and orientation of the contigs. Scaffolding by MeDuSa in this case increased the N50 value (from 184376 bp to 6963486 bp) by joining contigs together, putting most of the genome into a proposed order. This doesn't guarantee a better assembly but if the reference is an appropriate high quality genome assembly this may give more information on the location of genes in the genome. Figure 5.6B shows that much of the total genome length is in fewer than 4 scaffolds. Though, as can be seen in table 5.2 for MED001 and figure 5.6 for MED001-4 the genome(s) size is still considerably smaller than the reference, though the continuity of the assembly and N50 values have increased considerably. Particularly with a linear, high GC content genome which removes the possibility of a "closed" circular genome by possessing repetitive and hard to sequence telomeric regions as well as the complications of sequence accuracy that can arise due to high GC content. For the purposes of obtaining the best assembly possible before moving forward it would be beneficial for a more comprehensive reference genome to be obtained to provide a better scaffold. In addition, all de novo sequencing could be improved using PacBio sequencing and/or the inclusion of Sanger sequencing of telomeric regions.

Parameter	SPAdes Assembly	Scaffolding to *S. lavendulae*
# contigs/scaffolds (>= 0 bp)	285	219
# contigs/scaffolds (>= 1000 bp)	106	15
Total length (>= 1000 bp)	7058497	7071816
GC (%)	71.84	71.84
N50	184376	6963486

Table 5.2. Assembly statistics. Selected aspects of the QUAST report for MED001 before and after scaffolding to the reference genome for Streptomyces *lavendulae*

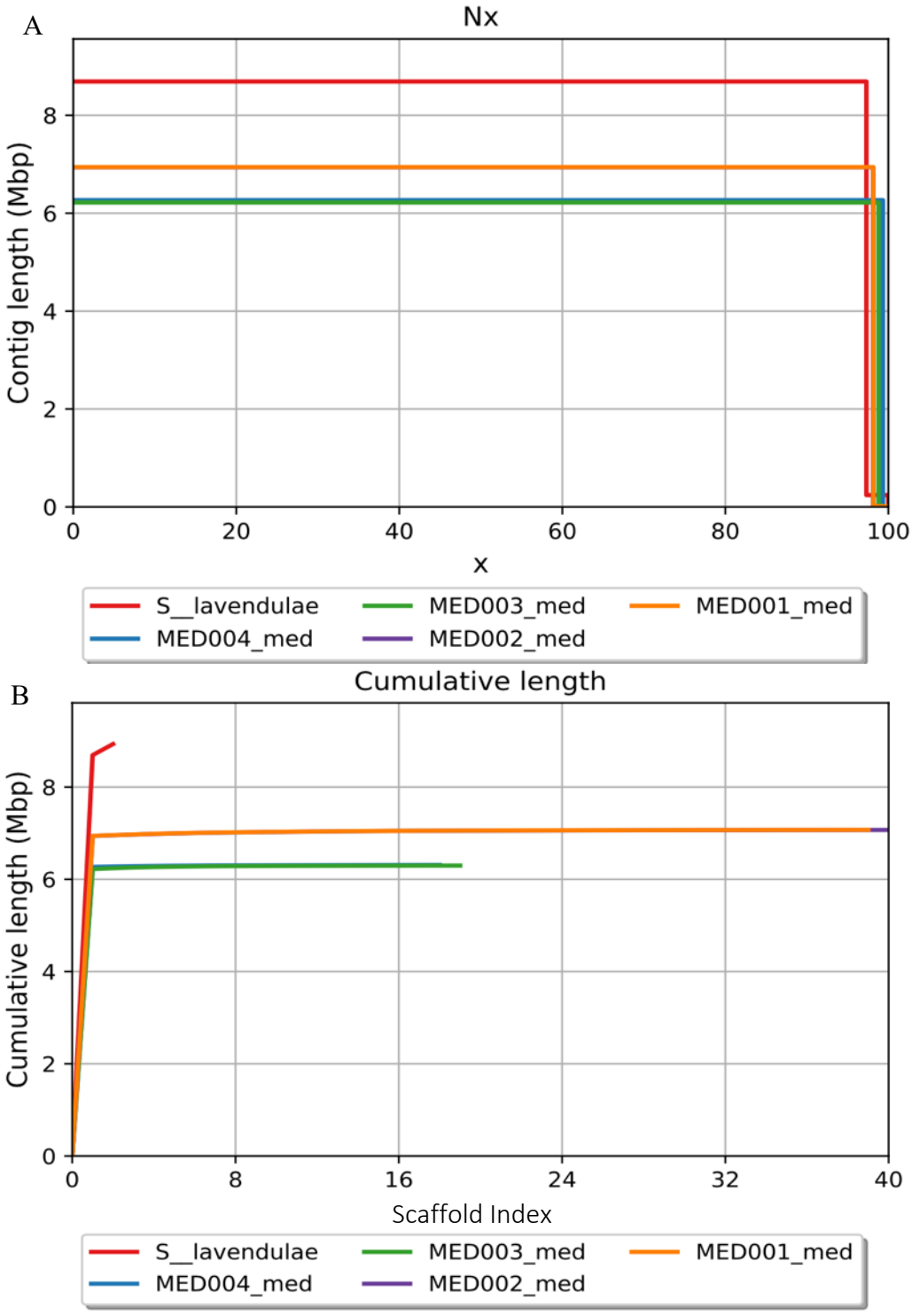


Figure 5.6. Assembly statistics graphs. Showing the assembly quality statistics as determined by QUAST of all four genomes after scaffolding with MeDuSa against *S. lavendulae* and the reference *S. lavendulae*. A) Displays the cumulative length of each assembly with the contig number on the X-axis. Only contigs of a size of 500 bp or larger were used here. B) gives the length of each contig in the assembly

5.2.4 Phylogenetic Placement of Strains of Interest Based on Sequencing Data

One potential reason for the discrepancy between the reference strain and the strains assembled here would be if the reference used was not the most appropriate. To determine this the phylogeny of the strains were evaluated using AutoMLST (Alanjary, Steinke and Ziemert, 2019). This had a secondary point of interest as the four genomes being analysed here are classified as being of three distinct strains which may indicate that the best assembly method and/or reference to use may not be universal across the data set. As such, finding out more about the phylogeny of all strains should help with the bioinformatic analysis. AutoMLST returned *Streptomyces albireticuli* as the most similar organism to all four of the strains (figure 5.7). Interestingly, this tree also puts the four strains being analysed here together, which suggested that at least within the boundaries of the genomes available to us at this time, these strains are more similar to each other than they are to anything else. The tree shows the four MED strains clustering in a node with only themselves and *S. albireticuli* whilst *S. lavendulae* is more distant which may explain why it didn't act as an ideal reference. Unfortunately, since there is no genome available for *S. ardus* or *S. caespitosus* these are not included in this analysis. Thus, since there is also a complete genome available for *Streptomyces albireticuli* it was used to scaffold the genome of MED001. As can be seen from table 5.3 this generated a more contiguous assembly than either of the previous methods. The results in table 5.3 show the genomes in >10 contigs, however, upon closer inspection of the N50 and the contig size graphs in Figures 5.8 and 5.9 it can be observed that almost all the genome is in one scaffold after reference guided assembly with MeDuSa. Whilst this is an improved assembly, the genomes are considerably shorter than either reference. This could be due to inadequate sequencing, improper assembly, or biological differences in the strains. It may also be that this reference still isn't the most appropriate, since AutoMLST uses sequence similarity over a number of genes selected based on genes in the query organism(s) that have identity found by HMMer in the 50 most similar genomes by average nucleotide identity (ANI) it is likely that there is not a more appropriate reference genome available. Before investigating the differences between the four MED strains and the reference further, the current genomes were polished to allow for the most accurate sequence-based comparison of the current genomes.

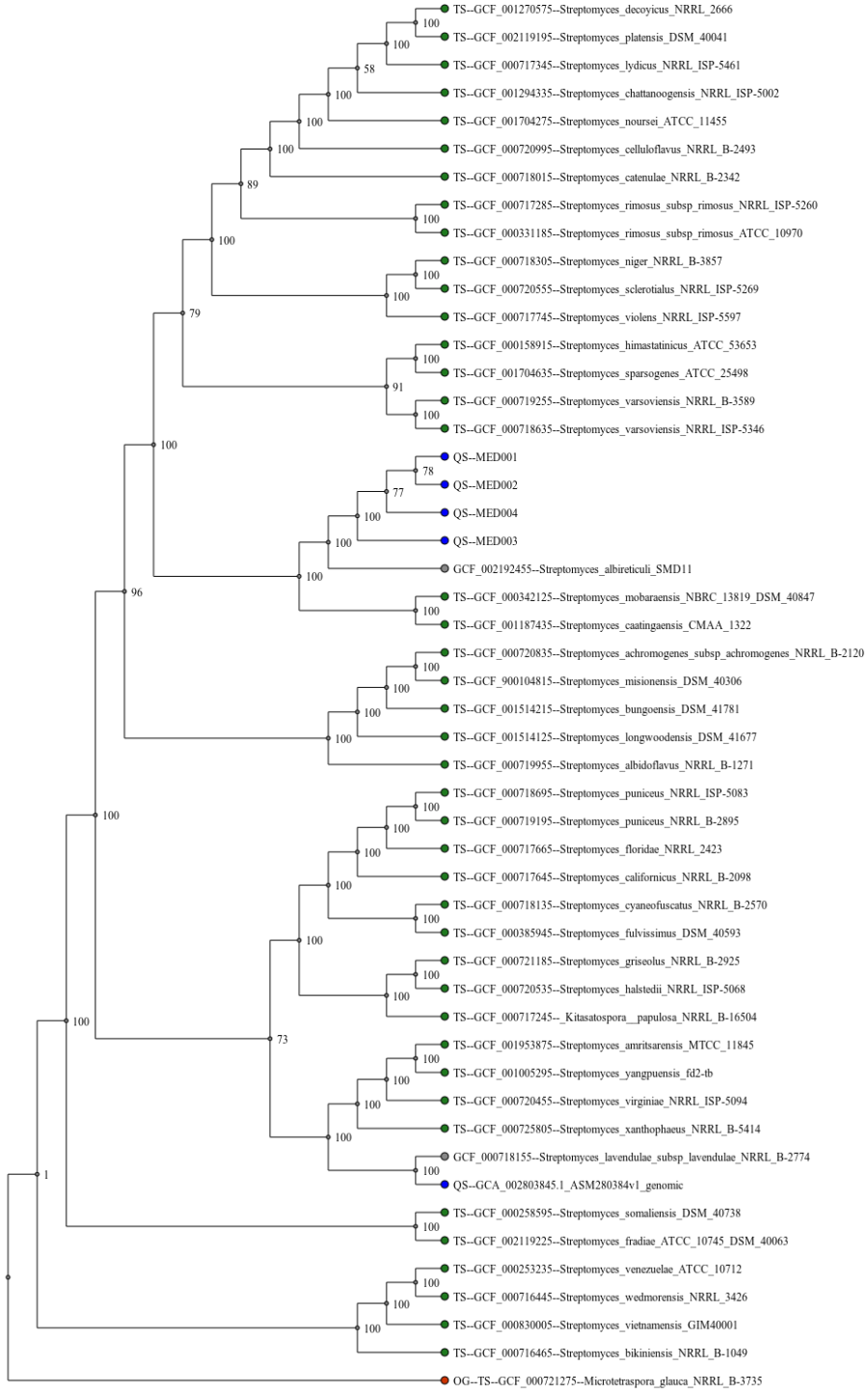


Figure 5.7. Phylogenetic Tree with MED strains. The concatenated alignment tree produced using AutoMLST with 100 bootstraps showing the four sequenced strains in a clade with *Streptomyces albireticul*i whilst *Streptomyces lavendulae* is more distant. The four strains used in this study are indicated by blue dots.

Parameter	SPAdes Assembly	Scaffolding to *S. lavendulae*	Scaffolded to *S. albireticuli*
# contigs / scaffolds (>= 0 bp)	285	219	209
# contigs / scaffolds (>= 1000 bp)	106	15	11
Total length (>= 1000 bp)	7058497	7071816	7077937
GC (%)	71.84	71.84	71.83
N50	184376	6963486	7000055

Table 5.3. Selected aspects of the QUAST report for MED001 before and after scaffolding to the reference genome for *Streptomyces lavendulae* and to *Streptomyces albireticul*i

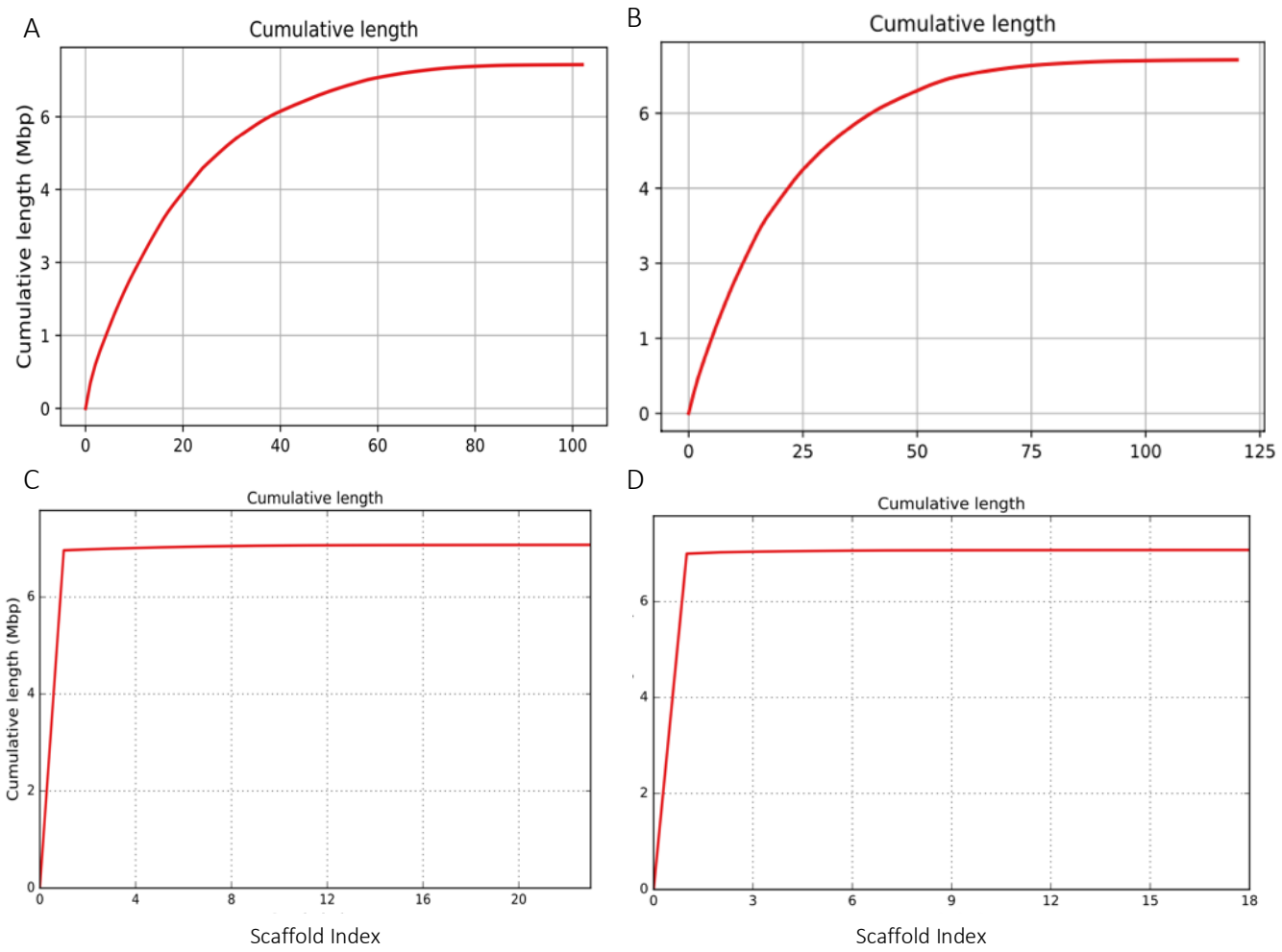


Figure 5.8. Assembly cumulative length plots. Data obtained by QUAST analysis of the different assembly methods. A. SPAdes assembly with careful correction. B. SPAdes assembly. C. SPAdes assembly with careful correction aligned against *Streptomyces lavendulae* into scaffolds using MeDuSa. D. SPAdes assembly with careful correction aligned against *Streptomyces albireticuli* into scaffolds using MeDuSa

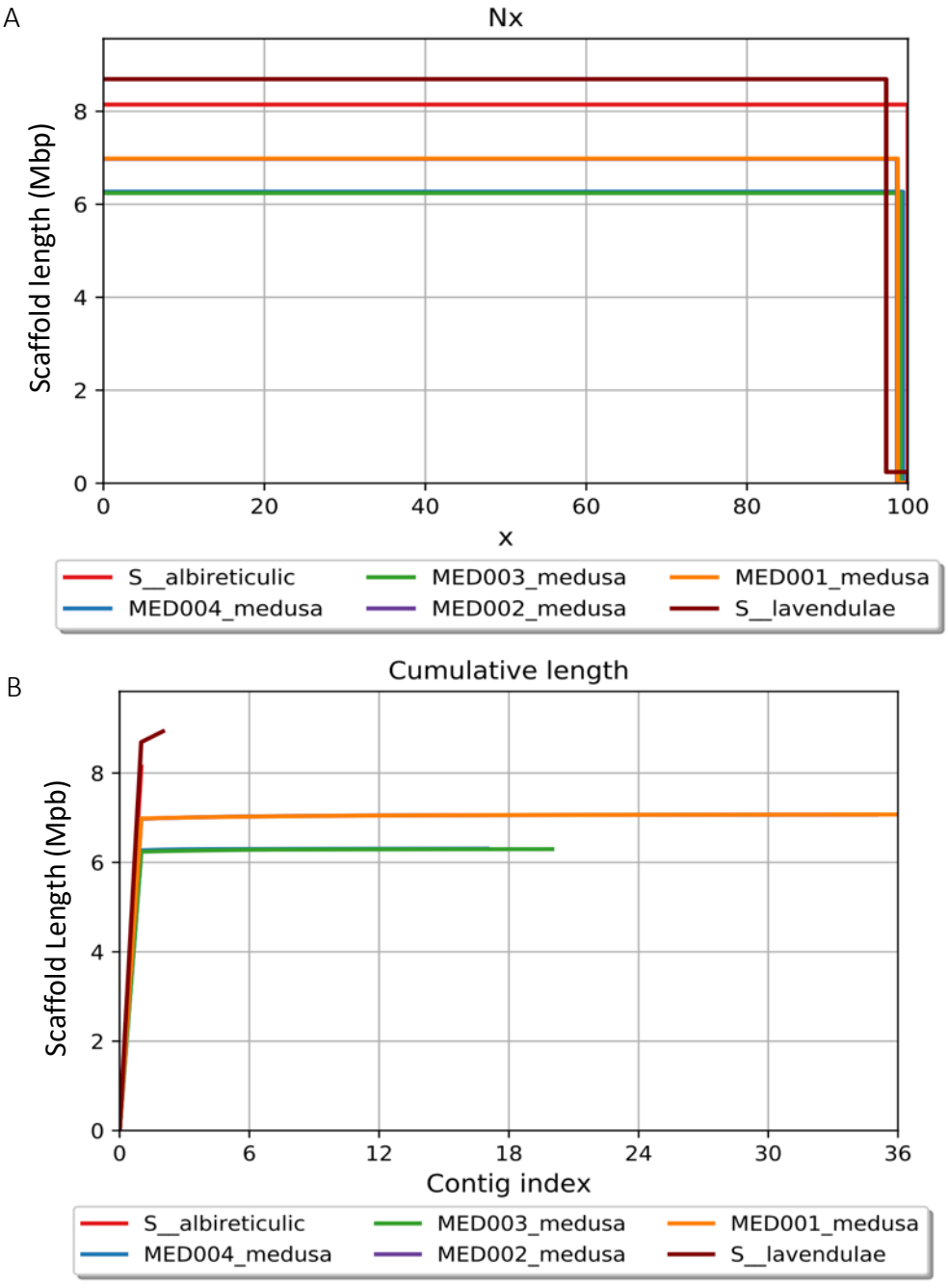


Figure 5.9. Assembly statistics as determined by QUAST Analysis. Data is given for all four genomes after scaffolding with MeDuSa against *S. albireticuli* and *S. albireticuliS. lavendulae* as the references for comparison A) Displays the cumulative length of each assembly with the contig number on the X-axis. Only contigs of a size of 500 bp or larger were used here. B) gives the length of each contig in the assembly

5.2.5 Assembly correction

After assembly scaffolding with MeDuSa, Pilon was used to correct the assembly, to check the effectiveness Pilon on these assemblies Snippy was run using the Illumina .fastq against the assembly of the same strain before and after Snippy. The results of this are shown in table 5.4 and suggest that whilst not 100% effective Pilon has improved a proportion of the errors in the assemblies.

Strain	SNPs before correction	SNPs after correction
MED001	5	4
MED002	32	10
MED003	26	2
MED004	5	5

Tables 5.4 Showing the number of SNPs per strain before and after assembly correction using Pilon. SNPs were determined using Snippy.

5.2.6 Assessment of Assembly Completeness

The assembly statistics shown indicate that the assembles for MED1-4 are of reasonable quality though it can be seen that the use of a loosely related strain for reference guided assembly has affected the final assembly. These metrics provide an indication of the completeness of the assembly of read but does not necessarily give us an accurate indication of how accurate the assembly is. To determine this BUSCO (Simão *et al.*, 2015), which uses the OrthoDB database of universally found single-copy orthologous genes to compare the number of Benchmarking Universal Single-Copy Orthologs (BUSCOs) found in each assembly to the number that would be expected in the genome based on evolutionary information. As can be seen in figure 6.10 the BUSCO assessment of the four assemblies suggest that they are almost entirely complete, with 351 out of 352 BUSCOs from the actinobacteria database searched for being found in a single copy. For comparison, both reference strains were also analysed using the same method, giving similar results. For *S. albireticuli* 3 BUSCOs were missing, 1 was a fragment and the remaining 348 were found as single copies. Similarly, for *S. lavendulae* 348 were found as single copies, 1 was found duplicated, 1 found fragmented and 2 were determined to be missing. Since both *S. albireticuli* and *S. lavendulae* are completed genomes, these results would suggest that all essential genes for the species are present in the current assemblies for MED001-04. Ideel analysis was also carried out to determine the number of proteins that are of the expected size when compared to the UniProt database. This indicated that whilst most proteins in these assemblies are of the expected size there are a considerable number that have been truncated. It is likely that the majority of these truncations are as the result of assemnly error rather than genuine biological differences. Given this, any anayisis based on gene truncation would require verification independent of these assemblies alone. Whilst wet lab work, such

as pulse field gel electrophoresis, would be required to show that these short genome assemblies are in fact truncated genomes and not misassemblies it would be unlikely that a misassembly would result in only non-essential genes being omitted from the assembly. As such it was taken that these assemblies could be used for further bioinformatic analysis in the meantime, with the view to verify this result in the lab if possible.

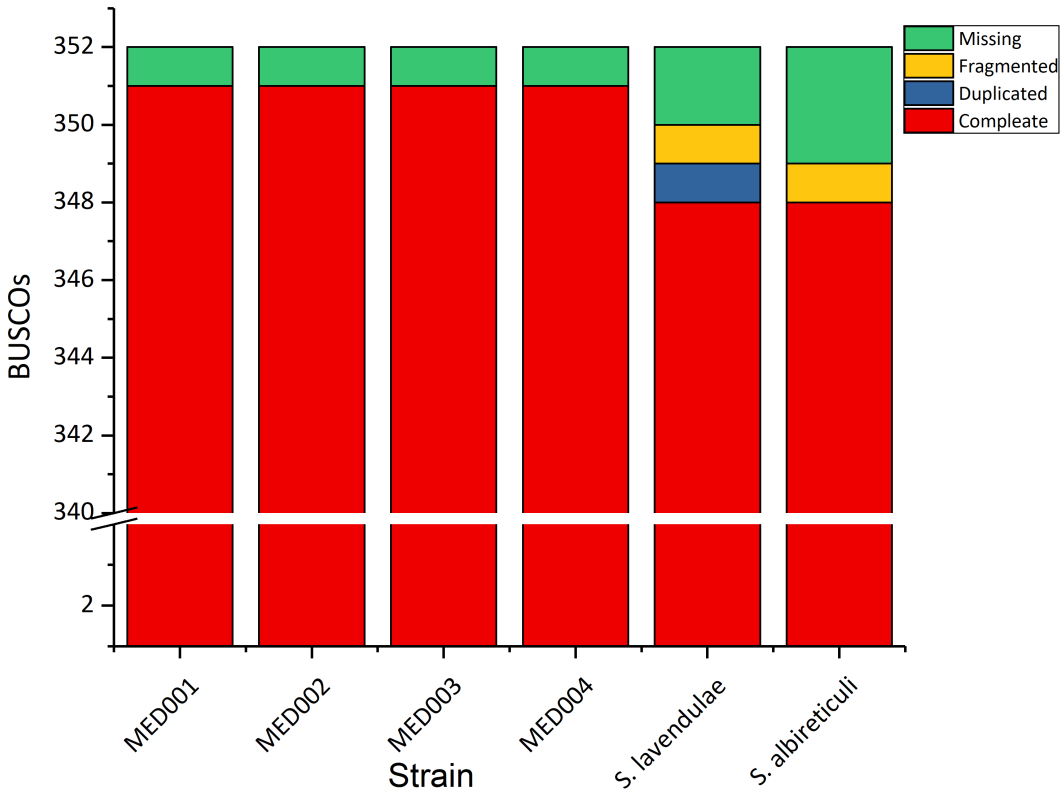


Figure 5.10 Summary of BUSCOs in assembled genomes. The figure shows the number of BUSCOs found to be complete (red), duplicated (blue), fragmented (yellow), or missing (green) in each of the four sequenced strains as well as two references.

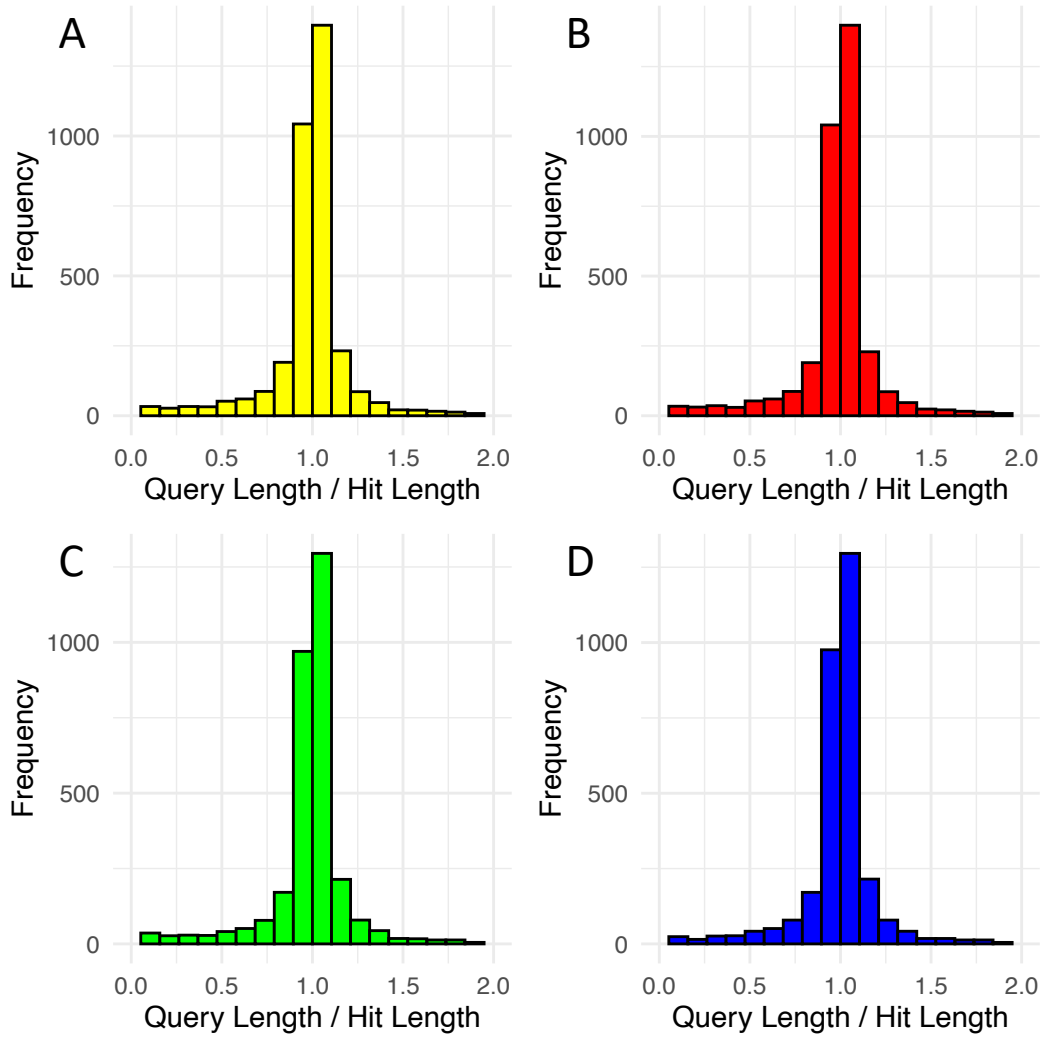


Figure 5.11. Ideel histogram. Output from Ideel in the form of histograms showing the relative sizes of predicted proteins as a ratio of the expected size in MED1-4 are shown in A-D respectively.

5.2.7 Alignment of Genomes with Reference

It has been noted that there are differences in the lengths of the genomes however, it has not been clear in the analysis done thus far whether this length discrepancy is due to one or a few larger deletions or systematic differences in the genomes. From previous results it is likely that the reduced length comes from various smaller deletions throughout the genomes rather than one or two large deletions. To see in more depth where this change occurs in the genomes whole-genome comparisons were made to visualise the genome architecture in relation to the reference using Mauve. The resulting alignment visualisation is shown in figure 5.12. Mauve shows areas that have a particular level of nucleotide similarity as a coloured block that corresponds to the same box in some location on the secondary genome given. In figure 5.12A MED001 in the lower panel has been aligned to *S. lavendulae* in the top panel, and whilst there are numerous sections where similarity is indicated each region is relatively short which suggests that the MED001 genome is potentially considerably fragmented or rearranged compared to *S. lavendulae.* Figure 5.12 B shows the same alignment from Mauve but with increased stringency on the requirement for similarity, with this the number of areas flagged as similar to a corresponding section on the other genome is minimal. These analyses suggest that it is more likely that the difference in size is due to the reference used being too phylogenetically distinct. Figure 5.13 shows all four MED strains aligned to one another, here significant regions of similarity can be observed but with obvious cases of rearrangement and/or inversion. At this stage it was hypothesised that *S. albireticuli* as a reference again may give more similarity. As such figure 5.14 A shows MED001 aligned to *S. albireticuli* which appears to show a similar level of rearrangement/dissimilarity as was seen with the alignment to *S. lavendulae.* Figure 5.13 B is the alignment of the same two strains but after the MED001 contigs were reordered by Mauve using *S. albireticuli* as a reference. In the reordered assembly there is still a clear difference in the two strains with movement of certain regions of interest in MED001 compared to the *S. albireticuli genome.* Finally, the MED002-4 contigs were reordered also by Mauve (figure 5.15) using MED001 as a reference, the regions of similarity here are much larger and with relatively similar positions in the genome architecture compared to when aligned with the reference. This arrangement may not be correct; however, it is better than random ordering of contigs and by having all four strains genomes arranged on the same basis further downstream analysis will be simpler. To determine the true genome architecture of the genomes a physical method would need to be employed.

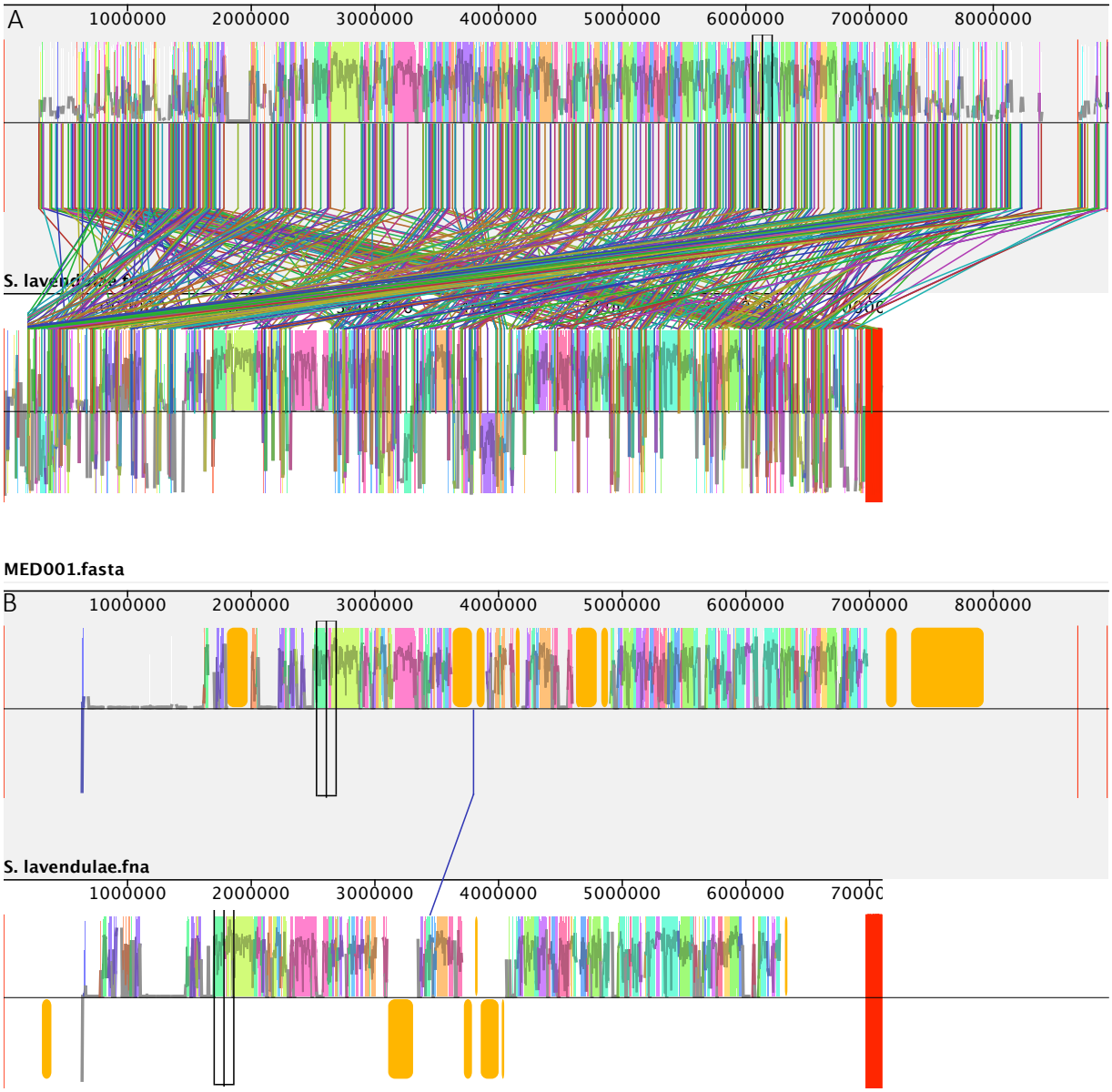


Figure 5.12. Mauve alignment visualisation of MED001 aligned to *S. lavendulae*. A) Shows the alignment with minimal LCB weights whilst B) shows the same alignment with maximum LCB weights.

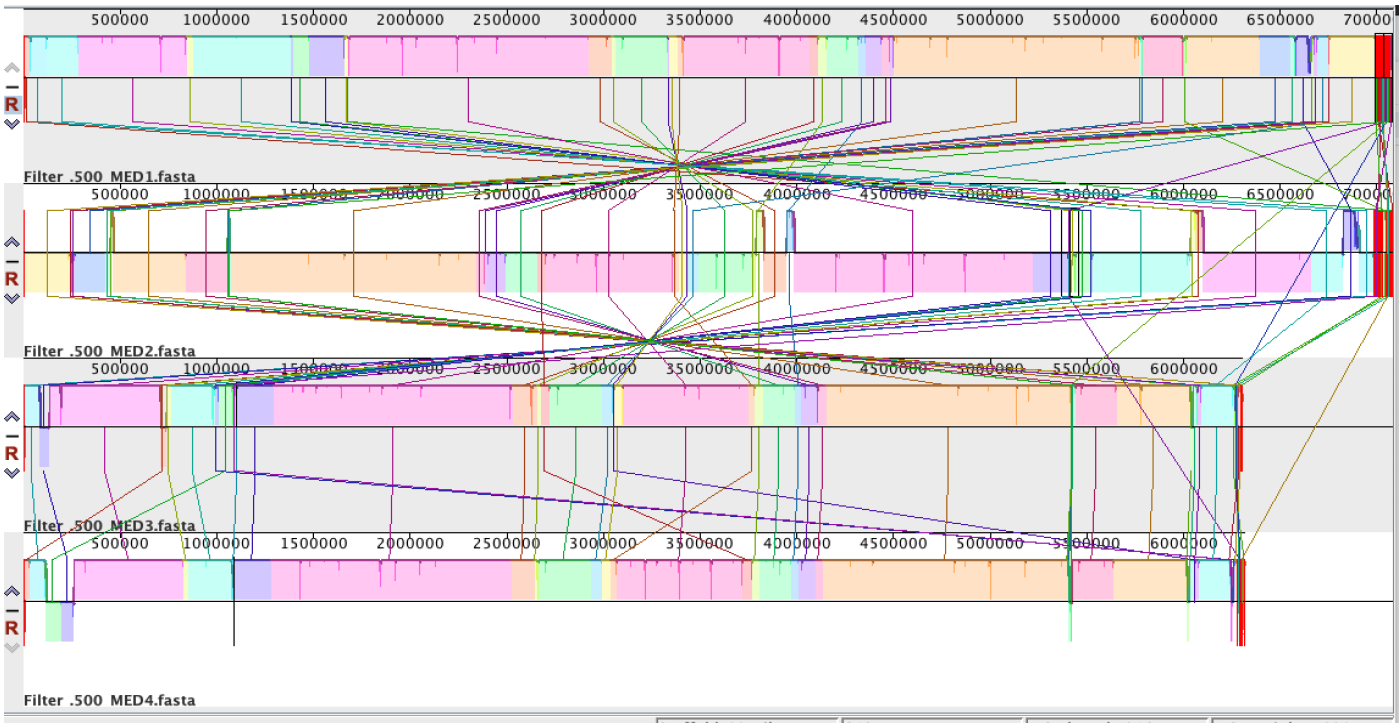


Figure 5.13 Mauve alignment. The four MED strains were aligned, where lines connect colour coded regions they are indicating regions of identity between strains.

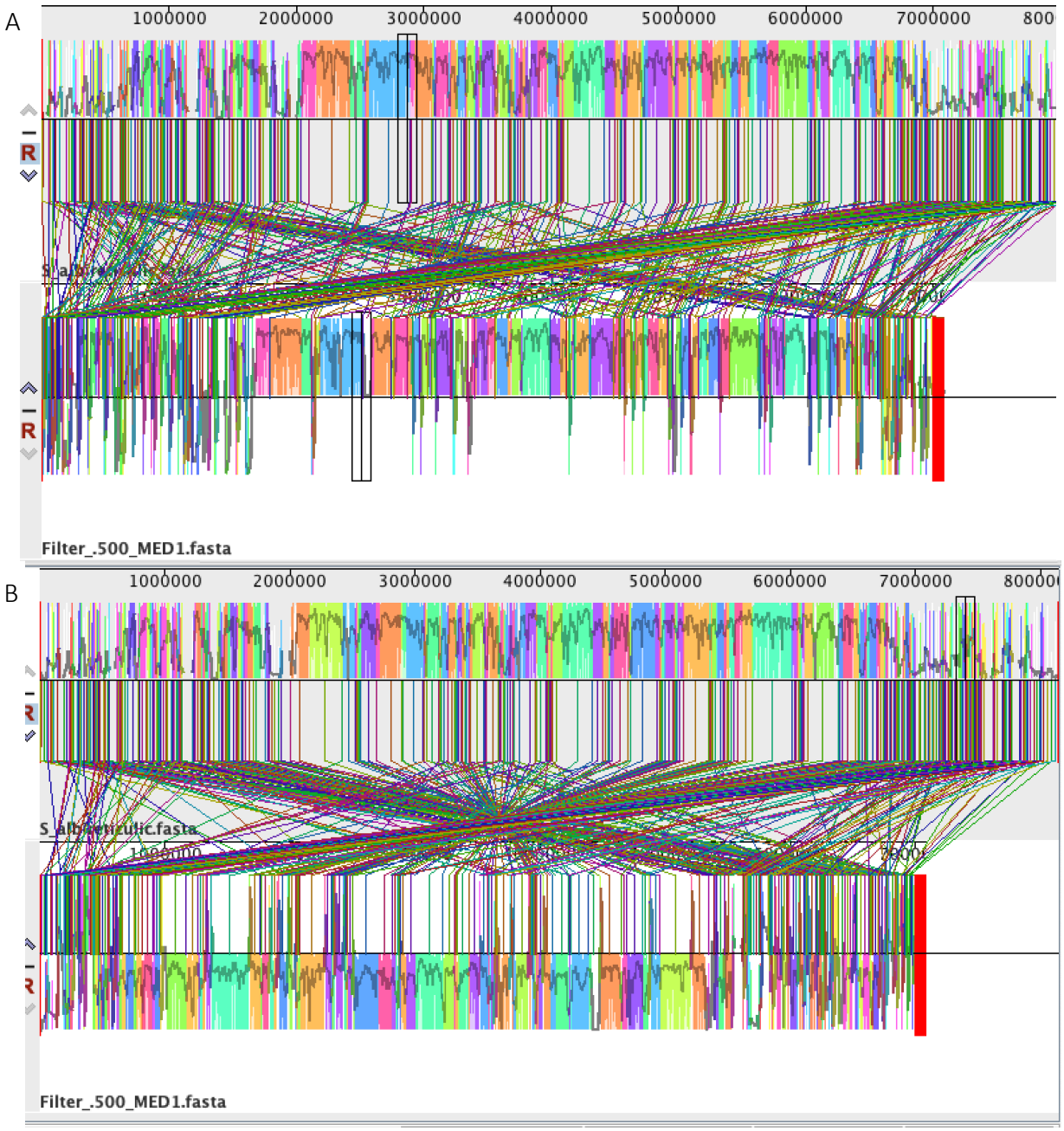


Figure 5.14 Mauve aligment. A) Mauve alignment of MED001 against *S. albireticuli* before contigs were moved B) the comparable alignment after moving of contigs by Mauve.

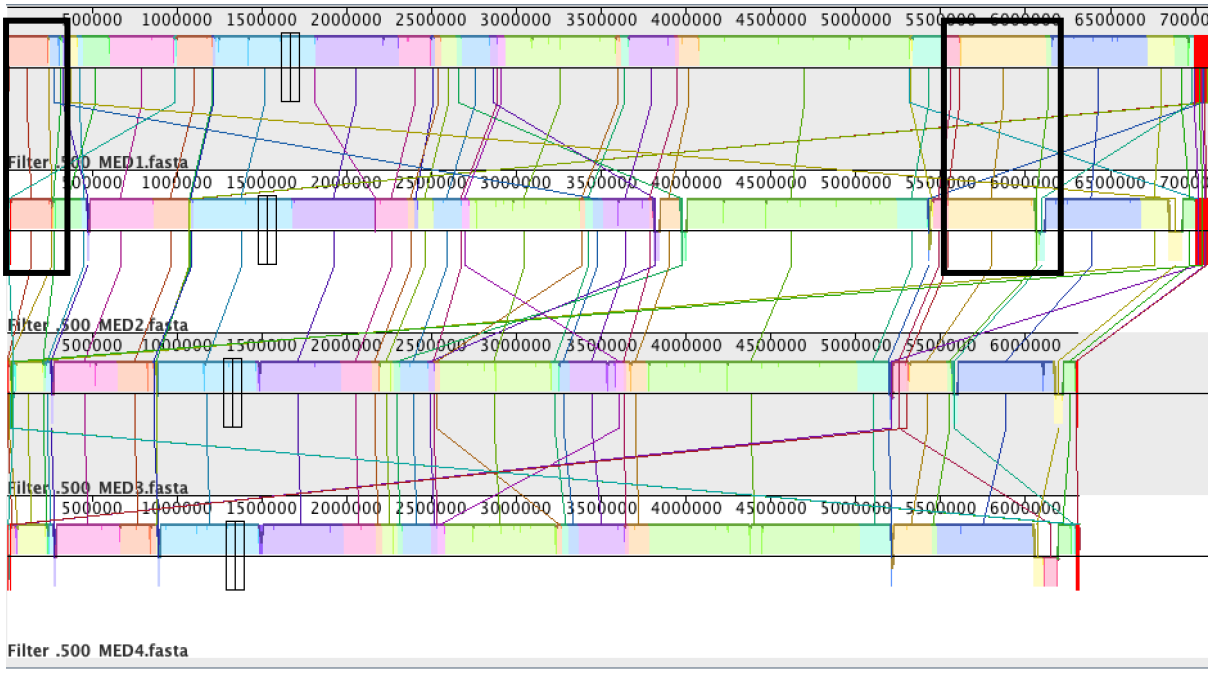


Figure 5.15. Mauve alignment of MED001, MED002, MED002 and MED004. Black boxes highlight regions of interest.

5.2.9 Comparison of Average Nucleotide Identity

FastANI (Jain *et al.*, 2018) was used to determine the average nucleotide identity between the four MED strains and the two references. The results are given in table 5.4 and figure 5.16.

	S. lavendulae	*S. albireticuli*	MED001	MED002	MED003	MED004
S. lavendulae	100	81.2063	81.0384	81.073	81.2006	81.2691
S. albireticuli	81.2128	100	86.28	86.2589	86.623	86.5821
MED001	80.972	86.3442	100	99.9958	99.965	99.9606
MED002	81.0082	86.315	99.9935	100	99.9513	99.952
MED003	81.2319	86.6304	99.9869	99.993	100	99.9897
MED004	81.2425	86.6739	99.9779	99.9895	99.9846	100

Table 5. 4 Average nucleotide identity. FastANI generated values for average nucleotide identity given as percentages.

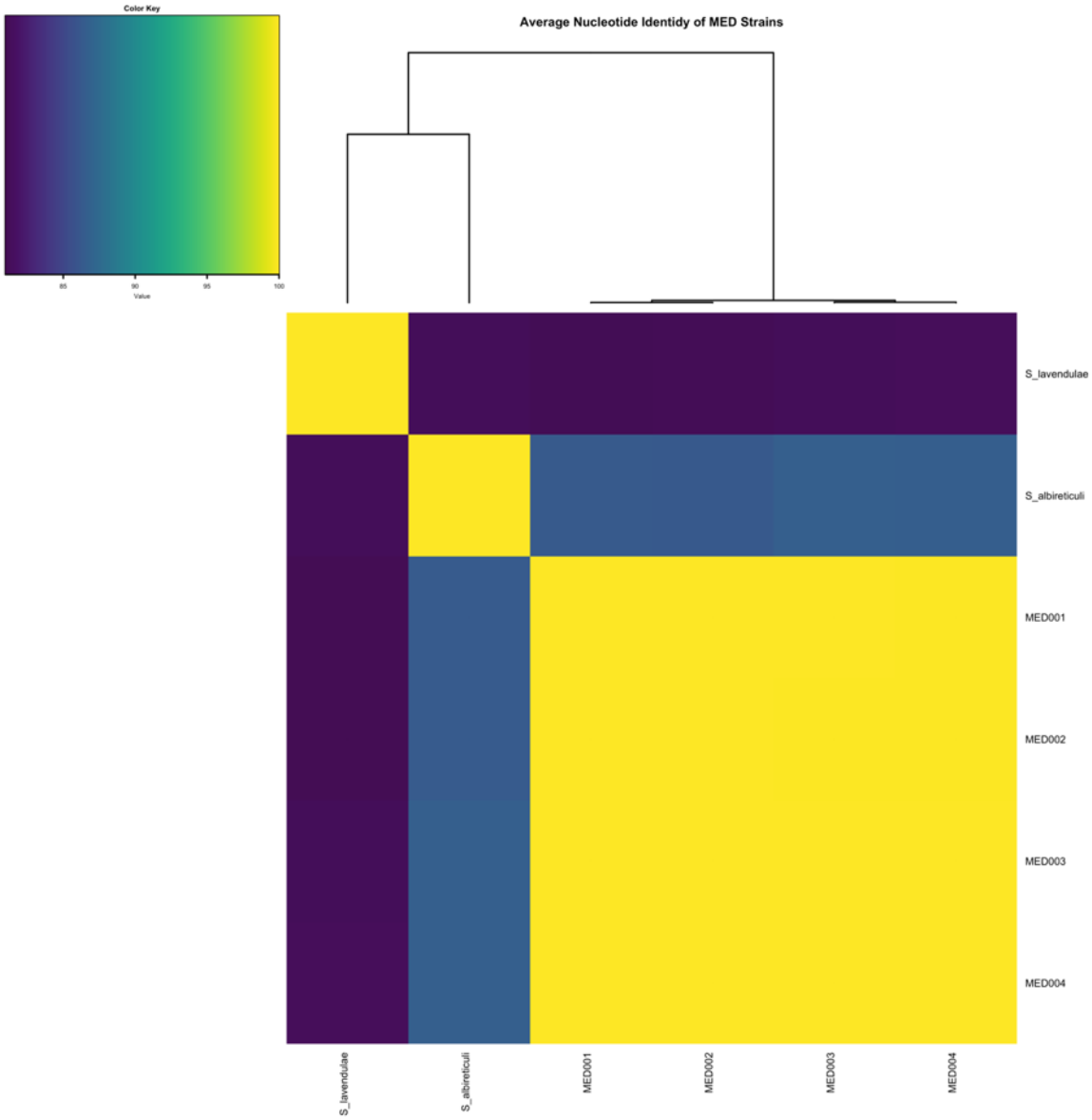


Figure 5. 16 Average nucleotide identity **heat map.** FastANI generated values for average nucleotide identity displayed as a heat map. Darker to lighter colours (Left to right on colour scale) indicates increasing identity level.

5.2.10 Comparison of Biosynthetic Potential

AntiSMASH 5.0 was used to analyse the biosynthetic potential of the four MED strains as well as *Streptomyces albireticuli* and *Streptomyces lavendulae* subsp. lavendulae (ASM280384v1) as reference. AntiSMASH was run using relaxed detection strictness to detect both partial and well defined clusters (Medema *et al.*, 2011; Blin *et al.*, 2013, 2017; Blin, Pascal Andreu, *et al.*, 2019; Blin, Shaw, *et al.*, 2019).This shows that there are fewer predicted biosynthetic clusters in MED003 and MED004 compared to MED001 and MED002, the initial hypothesis that one might draw from this information is that clusters that are costly in terms of carbon flux have been removed from the genomes of the more adapted strains to allow for increased production of a desired metabolite. This difference in the number of predicted clusters accounts for an average difference of 544,915 bp of predicted biosynthetic gene clusters between the early (MED001 and MED002) and late (MED003 and MED004) strains. As such the clusters "missing" from MED003 and MED004 where further investigated. From these predictions from antiSMASH there are a few key differences that can be seen, firstly the locations of most of the clusters follow generally a similar pattern between all strains indicated by the base numbers given for each predicted cluster in figures 5.17 – 5.20. This suggests that the efforts to make the assemblies as accurate as possible by using a reference guided approach to genome scaffolding and alignment based contig arrangements were successful. This is hypothesised that these artefacts may be because of difficulties to assembles the regions that contain some of the biosynthetic gene clusters as it can be seen that a disproportionate number of the clusters in MED002 are contained at the end of the genome. In total there are 4 predicted clusters that have a similarity of >20% identity to the cluster of a known metabolite that were found in the early strains but not the late. Most notably of these was geosmin, for which a biosynthetic gene cluster with 100% to the published geosmin cluster was found in MED001 and MED002 but not MED003 or MED004, the cluster as predicted from AntiSMASH is given in figure 5.16. Geosmin is one of the most well-known secondary metabolites produced by *Streptomyces* species as it is produced to such an extent that it produces a recognisable smell to humans, commonly associated with a 'post rain' smell. Geosmin, unlike many other *Streptomyces* natural products is not thought to have any effect on other microorganisms, is thought to be responsible for attracting spore spreading animals and deterring fruit flies. As is evidenced by the requirement and wide use of *Streptomyces* 'super-host' M1151 for improved heterologous expression of secondary metabolites, the production of not required metabolites in a production strain can divert valuable carbon flux away from processes required to produce the desired metabolites. Similarly, clusters with 100% similarity to the iso-migrastatin biosynthetic gene cluster from *Streptomyces platensis* subsp. *Rosaceus* (figure 5.17). Iso-migrastatin is known as a potent anti-cancer compound used clinically to inhibit the migration of tumour cells (Lim *et al.*, 2009).

Region	Type	From	To	Most similar known cluster	Type	Similarity
Region 1	NRPS-like,CDPS	21572	63795	albonoursin	Other	50%
Region 2	T1PKS	68367	119720	laidlomycin	Polyketide	50%
Region 3	T1PKS	156516	203842	herboxidiene	Polyketide	4%
Region 4	NAPAA,siderophore	225122	265222	desferrioxamine E	Other	100%
Region 5	T1PKS,hglE-KS,NRPS	276764	350015	sanglifehrin A	NRP + Polyketide	13%
Region 6	T1PKS	372824	422744	streptolydigin	type I + Saccharide:Hybrid/tailorin	15%
Region 7	terpene,lanthipeptide-class-iii	609304	636778	cepacin A	Polyketide	37%
Region 8	T1PKS,RRE-containing	650180	743288	sceliphrolactam	Polyketide	92%
Region 9	terpene,NAPAA	850719	906201	hopene	Terpene	76%
Region 10	NRPS,indole	1018478	1065813	formicamycins A-M	Polyketide	20%
Region 11	redox-cofactor	1089110	1111230	lankacidin C	NRP + Polyketide	20%
Region 12	ectoine,transAT-PKS	1249203	1334841	/ dorrigocin A / dorrigocin B / 13-epi-dorrigocin A	+ Polyketide:Trans-AT type I	100%
Region 13	hglE-KS,NRPS,T1PKS,indole	1339430	1423258	reductasporine	Other	66%
Region 14	T1PKS,other	1447595	1522098	mitomycin	Other:Aminocoumarin	70%
Region 15	T1PKS	1537136	1604318	mediomycin A	Polyketide	28%
Region 16	NRPS,T1PKS	1667971	1746840	antimycin	NRP + Polyketide	100%
Region 17	siderophore	1859566	1870684			0%
Region 18	terpene	2119534	2138825			0%
Region 19	melanin	3150647	3160611	melanin	Other	100%
Region 20	lassopeptide,T3PKS,NRPS	3344861	3421806	vazabitide A	NRP	60%
Region 21	thiopeptide,LAP	3462448	3492490	griselimycin	NRP	11%
Region 22	like,T2PKS,butyrolactone,T3PKS	3518906	3614131	granaticin	Polyketide:Type II	24%
Region 23	T2PKS,melanin	4780814	4853338	erdacin	Polyketide	28%
Region 24	lanthipeptide-class-iii	5187027	5209675	informatipeptin	RiPP:Lanthipeptide	42%
Region 25	PKS-like,NRPS	5356909	5414367	cyclooctatin	Terpene	100%
Region 26	ectoine	5622566	5632964	ectoine	Other	100%
Region 27	siderophore	5786610	5800304	goadsporin	RiPP:LAP	12%
Region 28	T1PKS	6127431	6171218	bafilomycin B1	Polyketide:Modular type I	50%
Region 29	RiPP-like	6215766	6224812			0%
Region 30	T3PKS,terpene	6297503	6338758			0%
Region 31	T3PKS	6483423	6524478	violapyrone B	Polyketide	28%
Region 32	T1PKS,NAPAA,LAP,thiopeptide	6552021	6639365	laidlomycin	Polyketide	56%
Region 33	terpene	6639923	6661913	geosmin	Terpene	100%
Region 34	lanthipeptide-class-iv	6666488	6689142			0%
Region 35	T1PKS	6713102	6769346	vazabitide A	NRP	19%
Region 36	like,lanthipeptide-class-i,terpene	6778706	6832937	paromomycin	Saccharide	7%
Region 37	NRPS-like,T1PKS	6838401	6886395	borrelidin	Polyketide:Modular type I	4%
Region 38	T1PKS,RiPP-like,terpene	6910094	6965337	jawsamycin	NRP + Polyketide	23%
Region 39	T1PKS,transAT-PKS	6971667	7084223	laidlomycin	Polyketide	37%

Figure 5.17. AntiSMASH Predicted Biosynthetic Gene Clusters. MED001 cluster location and similarity from AntiSMASH

Region	Type	From	To	Most similar known cluster	Type	Similarity
Region 1	NRPS-like,CDPS	48901	91124	albonoursin	Other	50%
Region 2	T1PKS	103262	150128	herboxidiene	Polyketide	2%
Region 3	NAPAA,siderophore	171408	211508	desferrioxamine E	Other	100%
Region 4	T1PKS,hglE-KS,NRPS	223050	296301	sanglifehrin A	NRP + Polyketide	13%
Region 5	T1PKS	318072	368980	streptolydigin	NRP + Polyketide:Modular t	15%
Region 6	terpene,lanthipeptide-class-iii	555540	583014	cepacin A	Polyketide	37%
Region 7	T1PKS,RRE-containing	596416	689524	sceliphrolactam	Polyketide	92%
Region 8	terpene,NAPAA	796949	852176	hopene	Terpene	76%
Region 9	NRPS,indole	933507	980842	formicamycins A-M	Polyketide	20%
Region 10	redox-cofactor	1004139	1026259	lankacidin C	NRP + Polyketide	20%
Region 11	ectoine,transAT-PKS	1164231	1249869	dorrigocin A / dorrigocin B / 13-epi-dorrigocin A	+ Polyketide:Trans-AT type I	100%
Region 12	hglE-KS,NRPS,T1PKS,indole	1254458	1338286	reductasporine	Other	66%
Region 13	T1PKS,other	1362623	1437081	mitomycin	Other:Aminocoumarin	70%
Region 14	NRPS,T1PKS	1554730	1633599	antimycin	NRP + Polyketide	100%
Region 15	T1PKS	1638281	1694845	piericidin A1	Polyketide:Modular type I	50%
Region 16	siderophore	1766758	1777876			0%
Region 17	terpene	2026356	2045928			0%
Region 18	melanin	3022795	3032759	melanin	Other	100%
Region 19	lassopeptide,T3PKS,NRPS	3176392	3255555	vazabitide A	NRP	60%
Region 20	thiopeptide,LAP,T2PKS,butyrolactone,T3PKS	3293562	3387230	granaticin	Polyketide:Type II	24%
Region 21	T2PKS,melanin	4628762	4701286	erdacin	Polyketide	28%
Region 22	lanthipeptide-class-iii	5034967	5057615	informatipeptin	RiPP:Lanthipeptide	42%
Region 23	PKS-like,NRPS	5204774	5262232	cyclooctatin	Terpene	100%
Region 24	ectoine	5470227	5480625	ectoine	Other	100%
Region 25	siderophore	5634415	5648109	goadsporin	RiPP:LAP	12%
Region 26	T1PKS	5972348	6033348	bafilomycin B1	Polyketide:Modular type I	55%
Region 27	RiPP-like	6077896	6086942			0%
Region 28	T3PKS,terpene	6159812	6201100			0%
Region 29	T3PKS,T1PKS,NAPAA,thiopeptide,LAP	6345765	6460735	laidlomycin	Polyketide	56%
Region 30	terpene	6461293	6483283	geosmin	Terpene	100%
Region 31	lanthipeptide-class-iv	6487858	6510512			0%
Region 32	T1PKS	6627010	6704583	laidlomycin	Polyketide	50%
Region 33	NRPS-like,T1PKS	6706012	6771881	vazabitide A	NRP	21%
Region 34	NRPS-like,lanthipeptide-class-i,terpene	6780375	6834606	paromomycin	Saccharide	7%
Region 35	NRPS-like,T1PKS	6840070	6888064	borrelidin	Polyketide:Modular type I	4%
Region 36	T1PKS,RiPP-like,terpene	6911763	6967006	jawsamycin	NRP + Polyketide	23%
Region 37	transAT-PKS,T1PKS	6972217	7085567	laidlomycin	Polyketide	37%

Figure 5.18. AntiSMASH Predicted Biosynthetic Gene Clusters. MED002 cluster location and similarity from AntiSMASH

Region	Type	From	To	Most similar known cluster	Type	Similarity
Region 1	T1PKS,NRPS,hglE-KS	3	102070	vazabitide A	NRP	23%
Region 2	siderophore,NAPAA	113612	153712	desferrioxamine E	Other	100%
Region 3	T1PKS	174992	239654	herboxidiene	Polyketide	2%
Region 4	T3PKS	275528	315034	violapyrone B	Polyketide	28%
Region 5	terpene,T3PKS	459826	501081			0%
Region 6	RiPP-like	573366	584190			0%
Region 7	T1PKS	627877	671015	bafilomycin B1	Polyketide:Modular type I	50%
Region 8	siderophore	999349	1012723	goadsporin	RiPP:LAP	12%
Region 9	ectoine	1167498	1177896	ectoine	Other	100%
Region 10	NRPS,PKS-like	1386088	1442567	cyclooctatin	Terpene	100%
Region 11	lanthipeptide-class-iii	1590557	1612353	informatipeptin	RiPP:Lanthipeptide	42%
Region 12	T2PKS,melanin	1946049	2017286	erdacin	Polyketide	28%
Region 13	T3PKS,T2PKS,butyrolactone,NRPS-like	3187861	3288704	granaticin	Polyketide:Type II	24%
Region 14	thiopeptide,LAP	3315079	3345886	griselimycin	NRP	11%
Region 15	NRPS,T3PKS,lassopeptide	3383169	3462332	vazabitide A	NRP	60%
Region 16	melanin	3652519	3662956	melanin	Other	100%
Region 17	terpene	4639938	4659369			0%
Region 18	siderophore	4907549	4919459			0%
Region 19	NRPS,T1PKS	5032122	5110991	antimycin	NRP + Polyketide	100%
Region 20	T1PKS	5174644	5233983	meridamycin	NRP + Polyketide	10%
Region 21	other,T1PKS	5247577	5323007	mitomycin	Other:Aminocoumarin	74%
Region 22	redox-cofactor	5381519	5403639	lankacidin C	NRP + Polyketide	20%
Region 23	NRPS,indole	5426850	5474352	formicamycins A-M	Polyketide	20%
Region 24	NAPAA,terpene	5585669	5642011	hopene	Terpene	76%
Region 25	T1PKS,RRE-containing	5749988	5843096	sceliphrolactam	Polyketide	92%
Region 26	lanthipeptide-class-iii,terpene	5855819	5884069	cepacin A	Polyketide	37%
Region 27	T1PKS	6069809	6123614	streptolydigin	NRP + Polyketide:Modular type I + Saccharide:Hybrid/tailoring	15%
Region 28	NRPS-like,CDPS	6241527	6284490	albonoursin	Other	50%

Figure 5.19. AntiSMASH Predicted Biosynthetic Gene Clusters. MED003 cluster location and similarity from AntiSMASH

Region	Type	From	To	Most similar known cluster	Type	Similarity
Region 1	T1PKS,NRPS,hglE-KS	19	115595	sceliphrolactam	Polyketide	24%
Region 2	siderophore,NAPAA	127137	167237	desferrioxamine E	Other	100%
Region 3	T1PKS	188517	253193	Sch-47554 / Sch-47555	Polyketide	3%
Region 4	T3PKS	291217	330723	violapyrone B	Polyketide	28%
Region 5	terpene,T3PKS	475515	516801			0%
Region 6	RiPP-like	588575	599399			0%
Region 7	T1PKS	643086	686902	bafilomycin B1	Polyketide:Modular type I	50%
Region 8	siderophore	1015029	1028403	goadsporin	RiPP:LAP	12%
Region 9	ectoine	1182958	1193356	ectoine	Other	100%
Region 10	NRPS,PKS-like	1402572	1459051	cyclooctatin	Terpene	100%
Region 11	lanthipeptide-class-iii	1607117	1628913	informatipeptin	RiPP:Lanthipeptide	42%
Region 12	T2PKS,melanin	1962927	2034164	erdacin	Polyketide	28%
Region 13	T3PKS,T2PKS,butyrolactone,NRPS-like	3254430	3346753	granaticin	Polyketide:Type II	21%
Region 14	thiopeptide,LAP	3373128	3403935	griselimycin	NRP	11%
Region 15	NRPS,T3PKS,lassopeptide	3441218	3520381	vazabitide A	NRP	60%
Region 16	melanin	3664533	3674970	melanin	Other	100%
Region 17	terpene	4651368	4670799			0%
Region 18	siderophore	4918754	4930664			0%
Region 19	NRPS,T1PKS	5043327	5122196	antimycin	NRP + Polyketide	100%
Region 20	T1PKS	5185849	5231038	s56-p1	NRP	11%
Region 21	redox-cofactor	5292129	5314249	lankacidin C	NRP + Polyketide	20%
Region 22	NRPS,indole	5337460	5384962	formicamycins A-M	Polyketide	20%
Region 23	NAPAA,terpene	5464962	5521304	hopene	Terpene	76%
Region 24	T1PKS,RRE-containing	5629281	5722389	sceliphrolactam	Polyketide	92%
Region 25	lanthipeptide-class-iii,terpene	5735113	5763363	cepacin A	Polyketide	37%
Region 26	T1PKS	5949826	5999979	streptolydigin	NRP + Polyketide:Modular type I + Saccharide:Hybrid/tailoring	15%
Region 27	other	6109568	6151856	mitomycin	Other:Aminocoumarin	67%
Region 28	NRPS-like,CDPS	6210202	6253165	albonoursin	Other	50%
Region 29	T1PKS	6283196	6319388	hygrocin A / hygrocin B	Polyketide	16%

Figure 5.20. AntiSMASH Predicted Biosynthetic Gene Clusters. MED004 cluster location and similarity from AntiSMASH

5.2.11 SNPs and Indel Detection

To detect the presence of single nucleotide polymorphisms (SNPs), insertions and deletions (Indels) Snippy (Galaxy Version 4.3.6+galaxy2) was used with the Illumina reads of each strain against the assembly of the comparison genome, MED002. Each strain was compared in this way to MED002 as patent information on the strains indicate that this is the best mitomycin C producer. The SNPs found are given in supplementary tables 1-3. Figure 5.21 illustrates the location of the SNPs found in MED001, MED003 and MED004 relative to MED002 which is marked with the locations of AntiSMASH predicted biosynthetic gene clusters. MED004 appears to be almost entirely SNPs, this is due to the resolution of this type of chat being that the size of one SNP marker is not proportional to the genome size shown here. This figure still serves to illustrate the relative abundance of single nucleotide polymorphisms to be greater in MED004 as well as where SNPs cluster on the genome and their relative localisation compared to AntiSMASH predicted biosynthetic gene clusters.

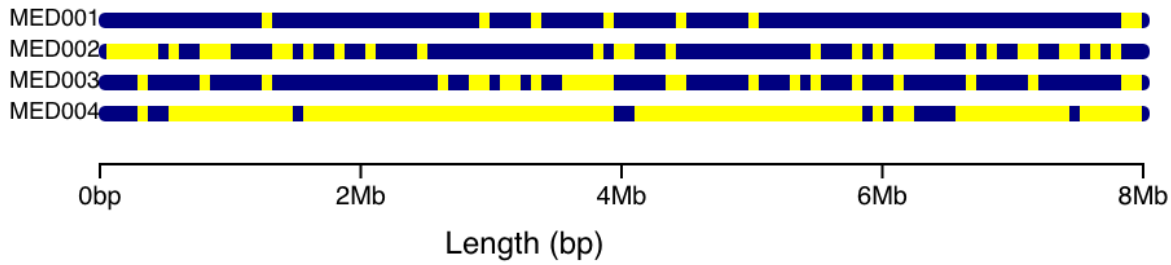


Figure 5.21. SNP location map. Plot generated with ChromoPlots in R showing the positioning of SNP results for predicted biosynthetic gene clusters in MED002 and the positions of SNPs present in MED001, MED003 and MED004 relative to MED002. The location of SNPs are indicated in yellow and are not to size scale.

5.2.12 Whole-genome Visualisation and Comparison

Given that the Mauve alignments displayed some differences between the early strains (MED001 and MED002) and the late strains (MED003 and MED004) CTT maps were used to look into these differences further. The initial map is given in 5.22 (DNA) and figure 5.23 (CDS) and shows all four MED strains as well as the two reference strains. In this map it can be clearly seen that MED001 and MED002 are very similar save for a few small regions of lower sequence identity. MED003 and MED004 are also still very similar to MED002 and whilst this doesn't show a direct comparison between MED003 and MED004 they also have a comparable level of similarity to MED001 suggesting that they are like each other. Interestingly the differences in these strains seems to be largely grouped in two main areas. Between 1 and 1.5 mbp and from 6 to 7 mbp there are regions of reduced sequence identity that also coincide with the locations of a significant proportion of the AntiSMASH predicted biosynthetic gene clusters.

Figure 5.24 gives more detail on the region of the mitomycin cluster in MED002. As can be seen there is only significant identity here to MED001 and not to the other two lineage strains or the reference genomes. This, however, does not explain the lack of mitomycin C production as it can be seen from the AntiSMASH results the part of the MED003 and MED004 genome that has identity to the mitomycin cluster is located at another part of these genome assemblies and as such it would not appear as high identity in this type of analysis. As such the ANTISmash predicted biosynthetic gene cluster from MED001 and MED004 was compared (figure 5.25) and it can be seen that there is an apparent truncation in the gene cluster. However, it is possible that the rest of the cluster exists in MED003 and MED004 but has been either misassembled or exists in a different chromosomal location than the latter half of the cluster. In order to determine which of these explanations was more likely. Sections of the beginning of the cluster in MED001 were searched for in MED004 by blast search. These sequences, including a non-ribosomal peptide synthetase were not found by Megablast search of the MED004 genome. This suggests that the truncated cluster is not due to an artefact of the genome assembly, it doesn't rule out that it is a result of sequencing sampling though this seems an unlikely coincidence for a region of the cluster of interest to have been undersampled for sequencing only in the two genomes known not to produce the compound of interest.

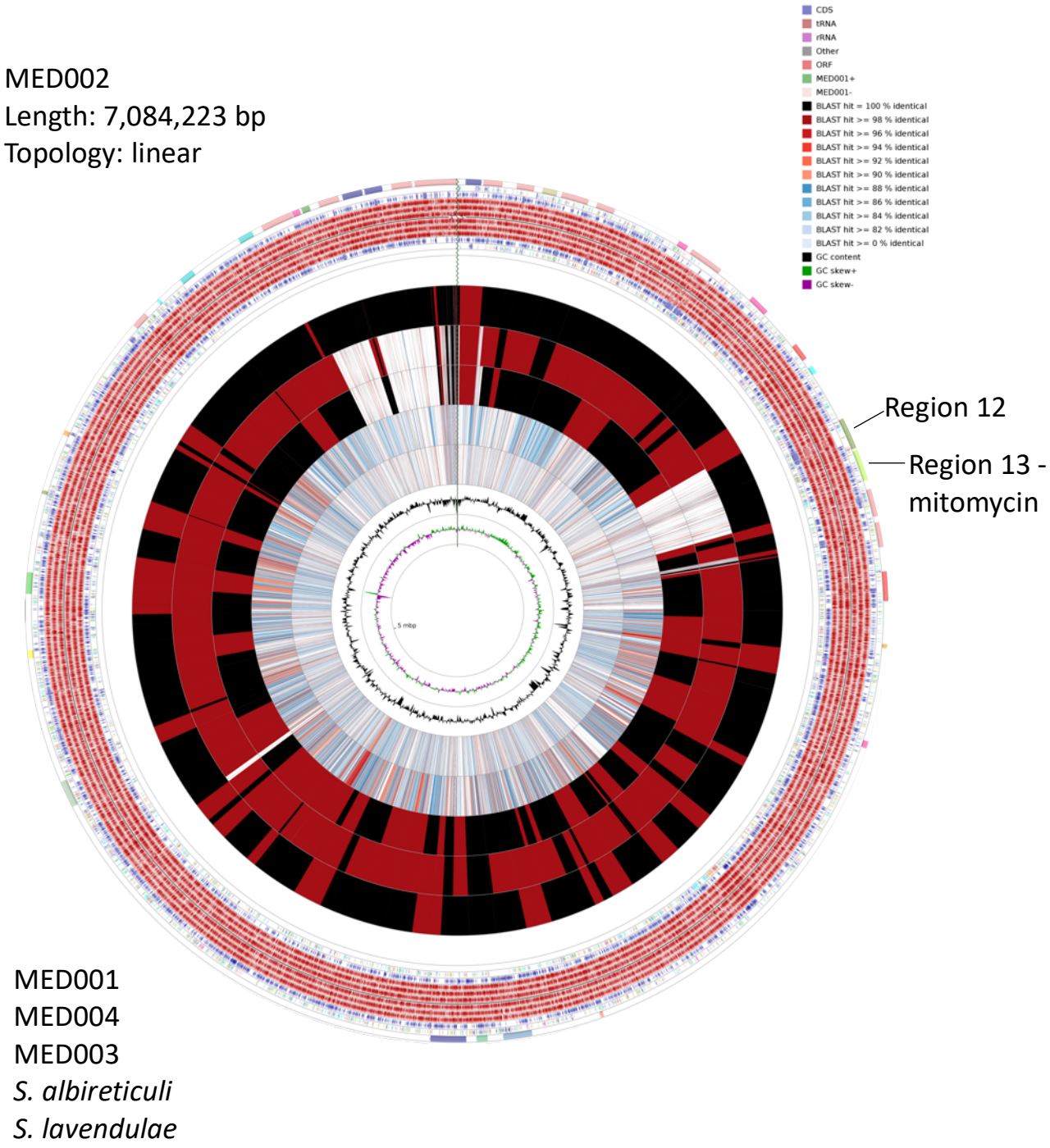


Figure 5.22 CCT map comparing the MED002 with the lineage and references. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. AntiSMASH predicted biosynthetic gene clusters. 2.COG functional categories for forward strand coding sequences; 3. Forward strand sequence features 4. Forward strand open reading frames; 5. Reverse strand open reading frames; 6. Reverse strand sequence features; 7. COG functional categories for reverse strand coding sequences. 8. the next 5 rings show regions of similarity based on DNA sequence identity between MED002and the comparison genomes in the order indicated in the figure. 9. GC content 10. GC skew

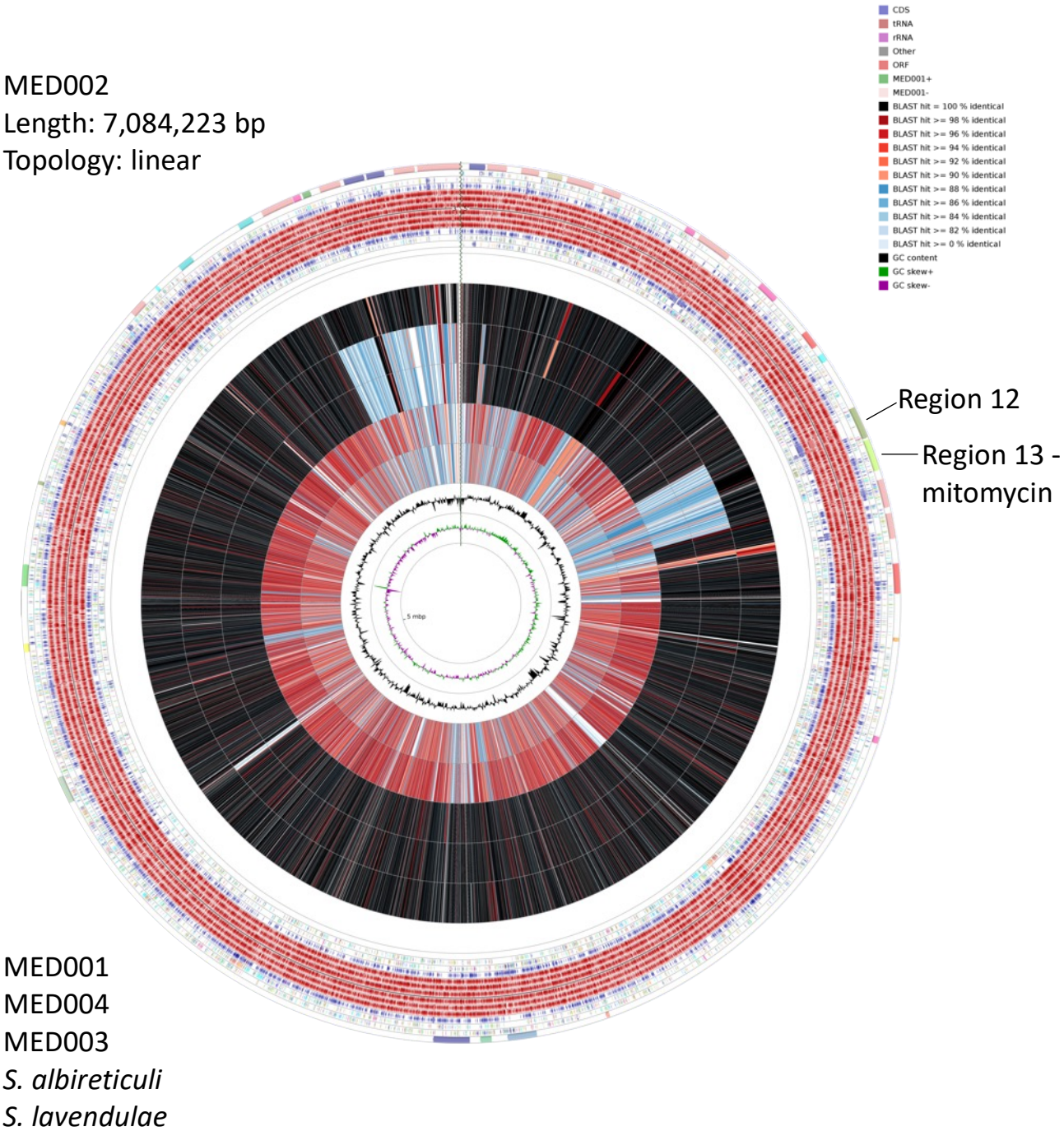
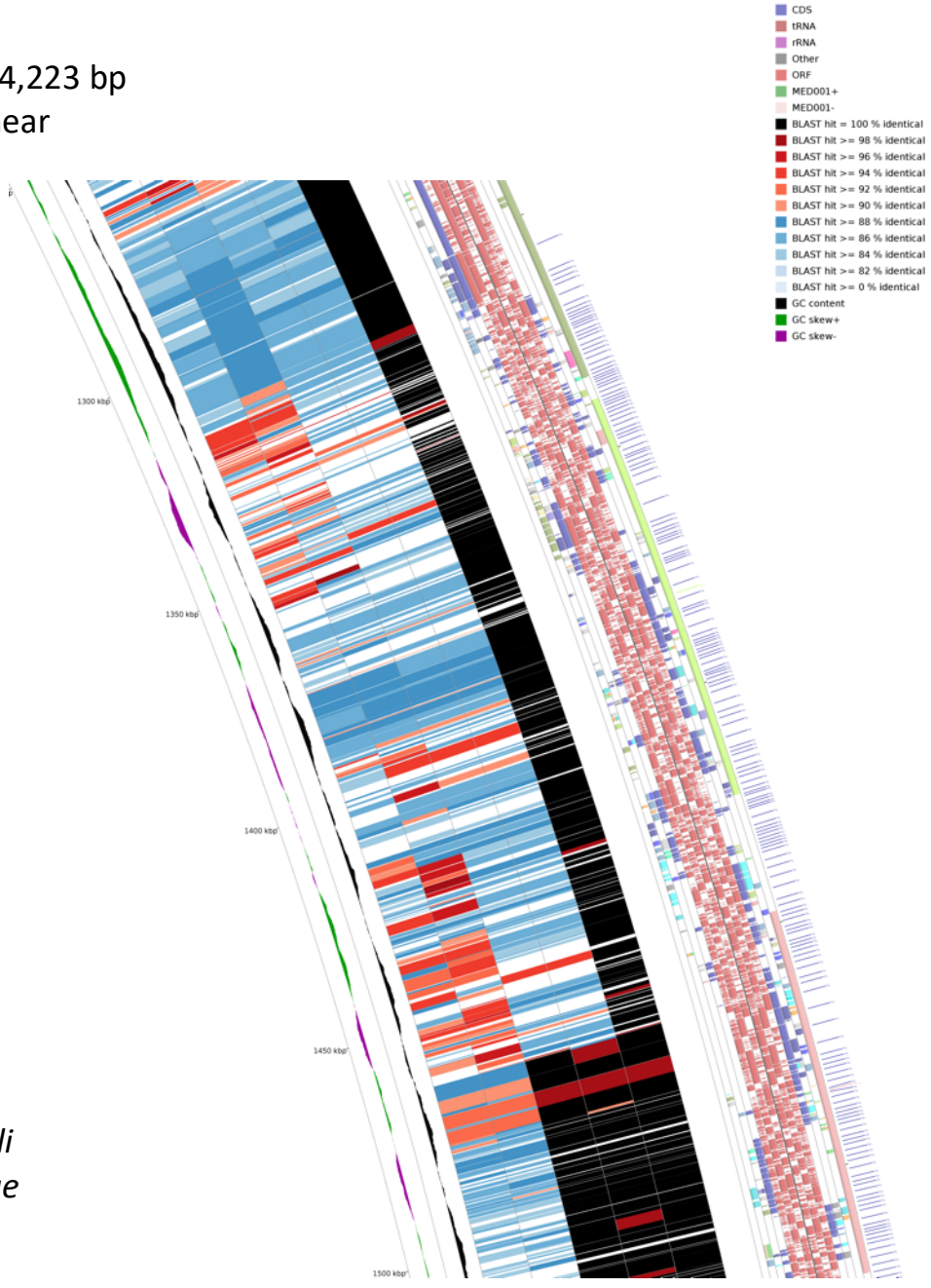


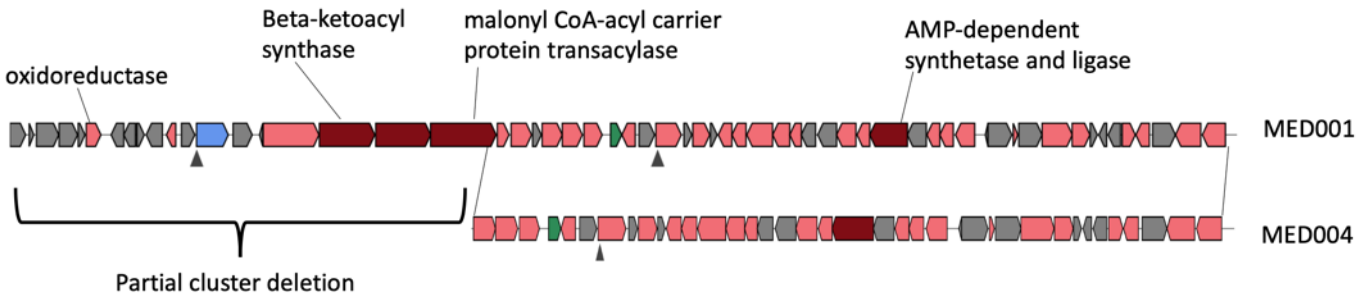
Figure 5.23 CCT map comparing the MED002 with the lineage and references. This map was generated using CCT maps build_blast_atlas.sh script. The map depicts from the outermost ring: 1. AntiSMASH predicted biosynthetic gene clusters. 2.COG functional categories for forward strand coding sequences; 3. Forward strand sequence features 4. Forward strand open reading frames; 5. Reverse strand open reading frames; 6. Reverse strand sequence features; 7. COG functional categories for reverse strand coding sequences. 8. the next 5 rings show regions of similarity as detected by BLAST comparison of CDS translations between MED002and the comparison genomes in the order indicated in the figure. 9. GC content 10. GC skew

MED002
Length: 7,084,223 bp
Topology: linear



MED001
MED004
MED003
S. albireticuli
S. lavendulae

5.24 Predicted mitomycin cluster comparison. The MED001 and MED004 clusters compared shows a considerable deletion, observed using the zoomed maps at 1399849 bp with 20 x zoom.



5.25 Predicted mitomycin cluster comparison. The MED001 and MED004 clusters compared shows a considerable deletion.

5.3 Conclusions and Future work

This data shows that the strains presented here likely represent the first genomes of the species as no other examples of the species predicted to be most closely related by AutoMLST analysis in the NCBI database. It is also seen that in terms of secondary metabolite production there is limited relation to phylogeny, as the main known producer of the compound of interest - *Streptomyces caespitosus* - is not included in this phylogenetic analysis due to its lack of published genome sequence. This is a well-studied phenomenon that secondary metabolite production capacity and phylogeny are unrelated due to the high rates of horizonal biosynthetic gene cluster transfer among Streptomyces species. We also see here that a potential answer to the questions posed by industry can be obtained from this genomic information and a well-informed experimental plan can be generated from this information.

Future work in this area would certainly begin with PCR confirmation of this aspect of the assemblies, particularly the contig boundaries and telomeric regions as well as the BCG of interest to allow for higher confidence in SNP calling in this region. Additionally, complementation of the whole cluster to from MED001/MED002 to MED003 and MED004 following by phenotypic analysis to determine if mitomycin C production was conferred by the presence of the entire cluster.

Chapter 6. Discussion and Conclusions

With improving sequencing technologies, whole-genome sequencing is being utilised in several industrial contexts to augment, replace and improve upon smaller scale techniques currently used. This is largely made feasible by the cost reduction in all types of sequencing in recent years (Kumar, Cowley and Davis, 2019) allowing for whole-genome data to be obtained in a rapid and cost-effective manner. The availability of a high-quality reference genomes constructed using both long and short-read sequencing technologies, allowing for both quality and reads to span gaps, can provide additional information for contig orientation and order when used for scaffolding. Additionally, a reference genome of high quality is required for any comparative genomics (Lischer and Shimizu, 2017). However, studies of this nature need to continue to build up a larger whole-genome reference base, as seen in this work the availability of reliable genome sequences in databases can still be a limiting factor. This work has shown that genomic data provides industry confirmation of previous genome editing attempts, and the potential to demonstrate their intellectual property. The provision of a high quality reference genome upon which an industrial lineage is based will be valuable for comparison to resulting strains as more improvements are made.

This work has shown that small scale, in-house, Illumina sequencing with low-density multiplexing leads to extremely high-quality data. The quality of data generated here is an improvement upon previously obtained in-house sequencing using Ion Torrent chemistry and data from outsourced Illumina data (this work, figure 3.1) up to the point of writing. This work has also shown that there are several benefits to long read technologies that help assemblies in terms of lower contig number and in the confidence given to assemblies where long reads span regions of interest. However, it has also shown that long read, error-prone sequencing increased CDS fragmentation in assemblies and that multiple methods of quality assembly must be utilised. The provision of a high quality, complete reference genome for Ethanol Red improved upon what was previously publicly available. This work has also shown the considerable improvements that this assembly has over the best publicly-available sequence. The differences these improvements to make the analysis of related data as well as the assembly of related strains was demonstrated. As such it should stand that others, initially within the industrial collaboration and eventually in the wider community, will be able to utilise this data to improve their own analysis. Furthermore, it was not previously known how well the particularly Illumina instrumentation, MiniSeq, would cope with such high GC content sequences as was required for *Streptomyces* sequencing. The data obtained from these particular experiments led to a number of other similar projects being undertaken. Aside from quality, one of the other benefits of in-house sequencing, as opposed to outsourcing, is the rapid turnaround time; in the case of the *Streptomyces* sequencing in this work, from spore stock to assembled sequence took only one working week. Additionally, the benefits to industry of being able to work within pre-existing non-disclosure agreements, meaning that work cannot be outsourced therefore whole-genome sequencing is uncommon, offers a significant advantage.

As well as the information provided and discussed in this work, the whole-genome sequencing data for the industrial strains in this work can be utilised by industry and our industrial partners to prove ownership of a particular strain. This is something that is difficult to achieve by sequence-independent mechanisms but since genetically-modified microorganisms are patentable this is vital to industry and collaboration (Carroll, 2015).

From both the phenotypic and genotypic data we see that these each have benefits individually but yield significantly more information combined. The utilisation of genomic data when analysing the phenotypic data gave the information more context and helped to avoid the generation of false conclusions as data could be obtained from multiple sources. From this combined information we can establish which strains are the best able currently to perform under specific industrial conditions. From the combined data we can tell what traits a well-performing strain has and if these can be transferred to other industrially relevant strains and if so, how easily this can be done. This means that multiple genes could potentially be combined into a singular strain which would be of great benefit to industry. Without this combinatorial data future experimentation of this nature would not be possible.

Shown in this work is the, complete genome of novel *S. cerevisiae* strain Ethanol Red, as well as the near-complete sequences of six related industrially relevant strains. We give a detailed outline of the genomic differences between strains as well as their phenotype in over 500 different conditions. Also demonstrated is the ease of use of sequencing technologies after the definition of clear methods and analysis pipelines which can be used with minimal modifications with several different organisms. We observed that high quality genomic data can be vital to the designing of site-specific heterologous gene insertions. Sequencing and analysis pipelines were also tested with a series of, high GC *Streptomyces*, which are typically difficult to sequence (Tilak *et al.*, 2018). These were found to be phylogenetically diverse from other known strains in the databases with results obtained from the genomic data . Further, it was established that there was a likely truncation in the mitomycin C biosynthetic gene cluster which was resulting in no production by half of the lineage.

Discussed here is an only subsection of the phenotypic and genomic data obtained, which is another great benefit to the techniques used here. Whilst these data sets were obtained with certain hypothesis in mind it is also true that other hypotheses that exist may be tested in the future and could be answered by the same data.

Chapter 7. Future work

This work presented a large volume of data, much of which lends itself to further investigation. A number of the suggested genomic explanations for phenotypic differences observed between Ethanol Red, Angel and S288c could be corroborated by gene mutagenesis and complementation to recover the phenotype. This would be of particular interest now that a strain has been chosen by our industrial partners to base their lineage on. Given that neither strain outperformed in all conditions, the information presented here could be used to confer Ethanol Red with the desirable trains from Angel that it doesn't already possess.

Completion of the MAL gene insertions into the *S. cerevisiae* lineage to test the hypothesis that the resistance cassette is inhibiting an improved phenotype from emerging from transformed strains compared to their wild type counterparts.

To complement the genome assemblies presented in this work, it would be interesting to carry out pulse-field gel electrophoresis in order to determine the accuracy of our assemblies. Particularly that of Ethanol Red which appears to be a complete genome. This should be easy to achieve as most of the experimental protocol for this has already been followed, with success except for some required PFGE condition optimisation to obtain a clear image.

Additionally, functional knockout/in and complementation experiments with the *Streptomyces* strain to confirm that conclusion that the missing part of the mitomycin cluster is responsible for the ability to produce mitomycin C would be illuminating.

Chapter 8. References

Abo, B. O. *et al.* (2019) 'Lignocellulosic biomass for bioethanol: An overview on pretreatment, hydrolysis and fermentation processes', *Reviews on Environmental Health*. doi: 10.1515/reveh-2018-0054.
Abraham, L. M. *et al.* (2006) 'Mitomycin: Clinical applications in ophthalmic practice', *Drugs*. doi: 10.2165/00003495-200666030-00005.
Agrawal, K. (2007) 'Dactinomycin', in *xPharm: The Comprehensive Pharmacology Reference*. doi: 10.1016/B978-008055232-3.61538-7.
Ahn, J. W. *et al.* (2012) 'A new Arctic Chlorella species for biodiesel production', *Bioresource Technology*. doi: 10.1016/j.biortech.2012.09.026.
Aínsa, J. A. (2000) 'Practical Streptomyces Genetics. T. Kieser, M. J. Bibb, M. J. Buttner, K. F. Chater, D. A. Hopwood', *International microbiology: official journal of the Spanish Society for Microbiology*.
Ajanovic, A. (2011) 'Biofuels versus food production: Does biofuels production increase food prices?', *Energy*. doi: 10.1016/j.energy.2010.05.019.
Alanjary, M., Steinke, K. and Ziemert, N. (2019) 'AutoMLST: An automated web server for generating multi-locus species trees highlighting natural product potential', *Nucleic Acids Research*. doi: 10.1093/nar/gkz282.
Alba-Lois, L. and Segal-kischinevzky, C. (2010) 'Yeast Fermentation and the Making of Beer and Wine The History of Beer and Wine Production', *Nature Education*. doi: 10.1038/455699a.
Albertin, W. and Marullo, P. (2012) 'Polyploidy in fungi: Evolution after whole-genome duplication', *Proceedings of the Royal Society B: Biological Sciences*. doi: 10.1098/rspb.2012.0434.
Alexander, W. G. (2018) 'A history of genome editing in Saccharomyces cerevisiae', *Yeast*. doi: 10.1002/yea.3300.
Alexander, W. G., Doering, D. T. and Hittinger, C. T. (2014) 'High-efficiency genome editing and allele replacement in prototrophic and wild strains of saccharomyces', *Genetics*. doi: 10.1534/genetics.114.170118.
Ambroset, C. *et al.* (2011) 'Deciphering the molecular basis of wine yeast fermentation traits using a combined genetic and genomic approach', *G3: Genes, Genomes, Genetics*. doi: 10.1534/g3.111.000422.
Andrews, S. (2014) *FastQC A Quality Control tool for High Throughput Sequence Data.* Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Aouida, M. *et al.* (2014) 'Activities and specificities of homodimeric TALENs in Saccharomyces cerevisiae', *Current Genetics*. doi: 10.1007/s00294-013-0412-z.
Araki, K. *et al.* (1995) 'Site-specific recombination of a transgene in fertilized eggs by transient expression of Cre recombinase', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.92.1.160.
Ascencio, D. *et al.* (2017) 'Increased rates of protein evolution and asymmetric deceleration after the whole-genome duplication in yeasts', *BMC Evolutionary Biology*, 17(1), p. 40. doi: 10.1186/s12862-017-0895-1.
Asgher, M. *et al.* (2010) 'Comparative study on two commercial strains of saccharomyces cerevisiae for optimum ethanol production on industrial scale', *Journal of Biomedicine and Biotechnology*. doi: 10.1155/2010/419586.
Aziz, R. K. *et al.* (2008) 'The RAST Server: Rapid annotations using subsystems technology',

BMC Genomics. doi: 10.1186/1471-2164-9-75.
Bachmann, R. *et al.* (2000) 'Using plants as plants', p. 93. Available at: https://link.gale.com/apps/doc/A62799338/AONE?u=ustrath&sid=AONE&xid=a34c83c8.
Baichwal, V. R. *et al.* (1983) 'Leucine biosynthesis in yeast - Identification of two genes (LEU4, LEU5) that affect α-Isopropylmalate synthase activity and evidence that LEU1 and LEU2 gene expression is controlled by α-Isopropylmalate and the product of a regulatory gene', *Current Genetics*. doi: 10.1007/BF00445877.
Banerjee, A. and Biswas, I. (2008) 'Markerless multiple-gene-deletion system for Streptococcus mutans', *Applied and Environmental Microbiology*. doi: 10.1128/AEM.02346-07.
Bankevich, A. *et al.* (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*. doi: 10.1089/cmb.2012.0021.
Barke, J. *et al.* (2010) 'A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant Acromyrmex octospinosus', *BMC Biology*. doi: 10.1186/1741-7007-8-109.
Bass, P. D. *et al.* (2013) 'Mitomycinoid alkaloids: Mechanism of action, biosynthesis, total syntheses, and synthetic approaches', *Chemical Reviews*. doi: 10.1021/cr3001059.
Beaton, A. *et al.* (2018) 'Community-led comparative genomic and phenotypic analysis of the aquaculture pathogen Pseudomonas baetica a390T sequenced by Ion semiconductor and Nanopore technologies', *FEMS Microbiology Letters*. doi: 10.1093/femsle/fny069.
Becerra-Rodríguez, C., Marsit, S. and Galeote, V. (2020) 'Diversity of Oligopeptide Transport in Yeast and Its Impact on Adaptation to Winemaking Conditions', *Frontiers in Genetics*. doi: 10.3389/fgene.2020.00602.
Beilharz, T. *et al.* (2003) 'Bipartite signals mediate subcellular targeting of tail-anchored membrane proteins in Saccharomyces cerevisiae', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M212725200.
Bekeova, C. *et al.* (2016) 'Characterisation of the genes involved in the biosynthesis and attachment of the aminodeoxysugar d-forosamine in the auricin gene cluster of Streptomyces aureofaciens CCM3239', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-015-7214-9.
Berlin, K. *et al.* (2015) 'Assembling large genomes with single-molecule sequencing and locality-sensitive hashing', *Nature Biotechnology*. doi: 10.1038/nbt.3238.
Biolog (2007) 'BIOLOG Phenotype MicroArrays Panels PM-M1 to PM-M14', *Biolog*.
Blin, K. *et al.* (2013) 'antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers.', *Nucleic acids research*. doi: 10.1093/nar/gkt449.
Blin, K. *et al.* (2017) 'AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification', *Nucleic Acids Research*. doi: 10.1093/nar/gkx319.
Blin, K., Shaw, S., *et al.* (2019) 'antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline', *Nucleic acids research*. doi: 10.1093/nar/gkz310.
Blin, K., Pascal Andreu, V., *et al.* (2019) 'The antiSMASH database version 2: A comprehensive resource on secondary metabolite biosynthetic gene clusters', *Nucleic Acids Research*. doi: 10.1093/nar/gky1060.
Bochner, B. R. (2009) 'Global phenotypic characterization of bacteria', *FEMS Microbiology Reviews*. doi: 10.1111/j.1574-6976.2008.00149.x.
Boles, E. and Hollenberg, C. P. (1997) 'The molecular genetics of hexose transport in yeasts', *FEMS Microbiology Reviews*. doi: 10.1016/S0168-6445(97)00052-1.
Bosi, E. *et al.* (2015) 'MeDuSa: A multi-draft based scaffolder', *Bioinformatics*. doi:

10.1093/bioinformatics/btv171.
Bowman, S. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome XIII', *Nature*. doi: 10.1038/387s090.
Brennan, L. and Owende, P. (2010) 'Biofuels from microalgae-A review of technologies for production, processing, and extractions of biofuels and co-products', *Renewable and Sustainable Energy Reviews*. doi: 10.1016/j.rser.2009.10.009.
Brinkman, E. K. *et al.* (2018) 'Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks', *Molecular Cell*. doi: 10.1016/j.molcel.2018.04.016.
Broder, A. Z. (2000) 'Identifying and filtering near-duplicate documents', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/3-540-45123-4_1.
Buijs, N. A., Siewers, V. and Nielsen, J. (2013) 'Advanced biofuel production by the yeast saccharomyces cerevisiae', *Current Opinion in Chemical Biology*. doi: 10.1016/j.cbpa.2013.03.036.
BURGERS, P. M. J. *et al.* (1990) 'Revised nomenclature for eukaryotic DNA polymerases', *European Journal of Biochemistry*. doi: 10.1111/j.1432-1033.1990.tb19165.x.
Bussey, H. *et al.* (1995) 'The nucleotide sequence of chromosome I from Saccharomyces cerevisiae.', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/PNAS.92.9.3809.
Bussey, H. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome XVI.', *Nature*. doi: 10.1038/387s078.
Butti, S. K. and Mohan, S. V. (2017) 'Autotrophic biorefinery: dawn of the gaseous carbon feedstock', *FEMS microbiology letters*. doi: 10.1093/femsle/fnx166.
Byrne, K. P. and Wolfe, K. H. (2005) 'The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species', *Genome Research*. doi: 10.1101/gr.3672305.
Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*. doi: 10.1186/1471-2105-10-421.
Cantarel, B. L. *et al.* (2008) 'MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes', *Genome Research*. doi: 10.1101/gr.6743907.
Carlson, M. and Botstein, D. (1982) 'Two differentially regulated mRNAs with different 5′ ends encode secreted and intracellular forms of yeast invertase', *Cell*. doi: 10.1016/0092-8674(82)90384-1.
Carlson, M., Osmond, B. C. and Botstein, D. (1981) 'Mutants of yeast defective in sucrose utilization.', *Genetics*.
Carroll, M. W. (2015) 'Sharing research data and intellectual property law: A primer', *PLoS Biology*. doi: 10.1371/journal.pbio.1002235.
Case, R. J. *et al.* (2007) 'Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies', *Applied and Environmental Microbiology*. doi: 10.1128/AEM.01177-06.
Challis, G. L. and Hopwood, D. A. (2003) 'Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1934677100.
Chan, L. L. *et al.* (2011) 'Direct concentration and viability measurement of yeast in corn mash using a novel imaging cytometry method', *Journal of Industrial Microbiology and Biotechnology*. doi: 10.1007/s10295-010-0890-7.
Charron, M. J. *et al.* (1989) 'Molecular evolution of the telomere-associated MAL loci of

Saccharomyces.', *Genetics*.
Chater, K. F. (2016) 'Recent advances in understanding streptomyces', *F1000Research*. doi: 10.12688/f1000research.9534.1.
Chen, H., Li, T. and Wang, Q. (2019) 'Ten years of algal biofuel and bioproducts: gains and pains', *Planta*. doi: 10.1007/s00425-018-3066-8.
Chin, C. S. *et al.* (2013) 'Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data', *Nature Methods*. doi: 10.1038/nmeth.2474.
Chisti, Y. (2008) 'Biodiesel from microalgae beats bioethanol', *Trends in Biotechnology*. doi: 10.1016/j.tibtech.2007.12.002.
Cho, S., Shin, J. and Cho, B.-K. (2018) 'Applications of CRISPR/Cas System to Bacterial Metabolic Engineering', *International journal of molecular sciences*. MDPI, 19(4), p. 1089. doi: 10.3390/ijms19041089.
Churcher, C. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome IX', *Nature*.
Clarridge, J. E. (2004) 'Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases', *Clinical Microbiology Reviews*. doi: 10.1128/CMR.17.4.840-862.2004.
Cobb, R. E., Wang, Y. and Zhao, H. (2015) 'High-Efficiency Multiplex Genome Editing of Streptomyces Species Using an Engineered CRISPR/Cas System', *ACS Synthetic Biology*. doi: 10.1021/sb500351f.
Cock, P. J. A. *et al.* (2015) 'NCBI BLAST+ integrated into Galaxy', *GigaScience*. doi: 10.1186/s13742-015-0080-7.
Colleaux, L. *et al.* (1988) 'Recognition and cleavage site of the intron-encoded omega transposase', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.85.16.6022.
Compeau, P. E. C., Pevzner, P. A. and Tesler, G. (2011) 'How to apply de Bruijn graphs to genome assembly', *Nature Biotechnology*. doi: 10.1038/nbt.2023.
Conant, G. C. and Wolfe, K. H. (2007) 'Increased glycolytic flux as an outcome of whole-genome duplication in yeast', *Molecular Systems Biology*. doi: 10.1038/msb4100170.
Connor, T. R. *et al.* (2016) 'CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community', *Microbial genomics*. doi: 10.1099/mgen.0.000086.
Cooper, A. A. and Stevens, T. H. (1996) 'Vps10p cycles between the late-Golgi and prevacuolar compartments in its function as the sorting receptor for multiple yeast vacuolar hydrolases', *Journal of Cell Biology*. doi: 10.1083/jcb.133.3.529.
Dahlin, L. R. *et al.* (2018) 'Down-selection and outdoor evaluation of novel, halotolerant algal strains for winter cultivation', *Frontiers in Plant Science*. doi: 10.3389/fpls.2018.01513.
Dalisay, D. S. *et al.* (2013) 'Marine Sediment-Derived Streptomyces Bacteria from British Columbia, Canada Are a Promising Microbiota Resource for the Discovery of Antimicrobial Natural Products', *PLoS ONE*. doi: 10.1371/journal.pone.0077078.
Darling, A. C. E. *et al.* (2004) 'Mauve: Multiple alignment of conserved genomic sequence with rearrangements', *Genome Research*. doi: 10.1101/gr.2289704.
DeBoer, C., Dietz, A. and Lummis, N. E. (1960) 'Porfiromycin, a new antibiotic. I. Discovery and biological activities.', *Antimicrobial Agents Annual.*, pp. 17–22.
Delcher, A. L. *et al.* (2007) 'Identifying bacterial genes and endosymbiont DNA with Glimmer', *Bioinformatics*. doi: 10.1093/bioinformatics/btm009.
AL Delcher S Kasif, O White and SL Salzberg, D. H. (1999) 'Improved microbial gene

identification with GLIMMER', *Nucleic Acids Research*.
Demain, A. L. (1999) 'Pharmaceutically active secondary metabolites of microorganisms', *Applied Microbiology and Biotechnology*. doi: 10.1007/s002530051546.
Demirbas, A. (2009) 'Biofuels securing the planet's future energy needs', *Energy Conversion and Management*. doi: 10.1016/j.enconman.2009.05.010.
Deng, M. De and Coleman, J. R. (1999) 'Ethanol synthesis by genetic engineering in cyanobacteria', *Applied and Environmental Microbiology*. doi: 10.1128/aem.65.2.523-528.1999.
Devantier, R., Pedersen, S. and Olsson, L. (2005) 'Characterization of very high gravity ethanol fermentation of corn mash. Effect of glucoamylase dosage, pre-saccharification and yeast strain', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-005-1902-9.
Dietrich, F. S. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome V.', *Nature*.
Dix, D. R. *et al.* (1994) 'The FET4 gene encodes the low affinity Fe(II) transport protein of Saccharomyces cerevisiae', *Journal of Biological Chemistry*. doi: 10.1016/s0021-9258(18)47163-3.
Dna, D. (2013) 'An Introduction to Next-Generation Sequencing Technology', *Illumina*. doi: Pub No. 770-2012-008.
Dong, N., Liu, L. and Shao, F. (2010) 'A bacterial effector targets host DH-PH domain RhoGEFs and antagonizes macrophage phagocytosis.', *The EMBO journal*, 29(8), pp. 1363–76. doi: 10.1038/emboj.2010.33.
Droop, A. P. (2016) 'Fqtools: An efficient software suite for modern FASTQ file manipulation', *Bioinformatics*. doi: 10.1093/bioinformatics/btw088.
Du, H. and Liang, C. (2019) 'Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads', *Nature Communications*. doi: 10.1038/s41467-019-13355-3.
Duina, A. A., Miller, M. E. and Keeney, J. B. (2014) 'Budding yeast for budding geneticists: A primer on the Saccharomyces cerevisiae model system', *Genetics*. doi: 10.1534/genetics.114.163188.
Dutilh, B. E. *et al.* (2013) 'Explaining microbial phenotypes on a genomic scale: Gwas for microbes', *Briefings in Functional Genomics*. doi: 10.1093/bfgp/elt008.
Duval, E. H. *et al.* (2010) 'Microarray karyotyping of maltose-fermenting Saccharomyces yeasts with differing maltotriose utilization profiles reveals copy number variation in genes involved in maltose and maltotriose utilization', *Journal of Applied Microbiology*. doi: 10.1111/j.1365-2672.2009.04656.x.
Eddy, S. R. (1996) 'Hidden Markov models', *Current Opinion in Structural Biology*. doi: 10.1016/S0959-440X(96)80056-X.
Edgar, R. C. (2004) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*. doi: 10.1093/nar/gkh340.
Eisentraut, A. (2010) 'Sustainable Production of Second- generation Biofules -', *IEA Energy Papers*. doi: 10.1787/9789264084247-en.
Engel, S. R. *et al.* (2014) 'The reference genome sequence of Saccharomyces cerevisiae: then and now.', *G3: Genes, Genomes, Genetics*. doi: 10.1534/g3.113.008995.
Erickson, B., Lutt, E. and Winters, P. (2016) 'Can Biofuels Replace Fossil Fuels?', in *Consequences of Microbial Interactions with Hydrocarbons, Oils, and Lipids: Production of Fuels and Chemicals*. doi: 10.1007/978-3-319-31421-1_379-1.
Escalante-Chong, R. *et al.* (2015) 'Galactose metabolic genes in yeast respond to a ratio of

galactose and glucose', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1418058112.
Escobar-Henriques, M. and Daignan-Fornier, B. (2001) 'Transcriptional regulation of the yeast GMP synthesis pathway by its end products', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M007926200.
EW, J. and GR, F. (1982) 'Regulation of amino acid and nucleotide biosynthesis in yeast', in, pp. 181–299.
Fanning, E., Knippers, R. and Winnacker, E. L. (2012) *DNA Replication and the Cell Cycle: 43. Colloquium der Gesellschaft für Biologische Chemie, 9.–11. April 1992 in Mosbach/Baden*. Springer Berlin Heidelberg (Colloquium der Gesellschaft für Biologische Chemie in Mosbach Baden). Available at: https://books.google.co.uk/books?id=cpjzCAAAQBAJ.
Fatma, S. *et al.* (2018) 'Lignocellulosic Biomass: A Sustainable Bioenergy Source for Future', *Protein & Peptide Letters*. doi: 10.2174/0929866525666180122144504.
Feldmann, H. *et al.* (1994) 'Complete DNA sequence of yeast chromosome II', *EMBO Journal*. doi: 10.1002/j.1460-2075.1994.tb06923.x.
Fisk, D. G. *et al.* (2006) 'Saccharomyces cerevisiae S288C genome annotation: A working hypothesis', *Yeast*. doi: 10.1002/yea.1400.
Fogel, S. and Welch, J. W. (1982) 'Tandem gene amplification mediates copper resistance in yeast', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.79.17.5342.
Formighieri, C., Franck, F. and Bassi, R. (2012) 'Regulation of the pigment optical density of an algal cell: Filling the gap between photosynthetic productivity in the laboratory and in mass culture', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2012.02.021.
Förstemann, K. and Lingner, J. (2001) 'Molecular Basis for Telomere Repeat Divergence in Budding Yeast', *Molecular and Cellular Biology*. doi: 10.1128/mcb.21.21.7277-7286.2001.
Foury, F. (1997) 'Human genetic diseases: A cross-talk between man and yeast', *Gene*. doi: 10.1016/S0378-1119(97)00140-6.
Francis, S. (2010) 'Algal biofuels', in *Green Energy: Technology, Economics and Policy*. doi: 10.1201/b10163.
Freese, N. H., Norris, D. C. and Loraine, A. E. (2016) 'Integrated genome browser: Visual analytics platform for genomics', *Bioinformatics*. doi: 10.1093/bioinformatics/btw069.
Fujimura, T. *et al.* (1992) 'Pol of gag-pol fusion protein required for encapsidation of viral RNA of yeast L-A virus', *Nature*. doi: 10.1038/359746a0.
Gaffney, J. *et al.* (2020) 'Open access to genetic sequence data maximizes value to scientists, farmers, and society', *Global Food Security*. doi: 10.1016/j.gfs.2020.100411.
Galibert, F. *et al.* (1996) 'Complete nucleotide sequence of Saccharomyces cerevisiae chromosome X', *EMBO J*.
Gascón, S., NP, N. and JO, L. (1968) 'Comparative study of the properties of the purified internal and external invertases from yeast.', *The Journal of biological chemistry*, 243(7), pp. 1573–1577.
Georgieva, T. I., Skiadas, I. V. and Ahring, B. K. (2007) 'Effect of temperature on ethanol tolerance of a thermophilic anaerobic ethanol producer Thermoanaerobacter A10: Modeling and simulation', *Biotechnology and Bioengineering*. doi: 10.1002/bit.21536.
Gibson, B. *et al.* (2017) 'New yeasts-new brews: Modern approaches to brewing yeast design and development', *FEMS Yeast Research*. doi: 10.1093/femsyr/fox038.
Goecks, J. *et al.* (2010) 'Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences', *Genome Biology*.

doi: 10.1186/gb-2010-11-8-r86.
Goffeau, A. *et al.* (1996) 'Life with 6000 genes', *Science*. doi: 10.1126/science.274.5287.546.
Goldstein, S. *et al.* (2019) 'Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing', *BMC Genomics*. doi: 10.1186/s12864-018-5381-7.
Gomez-Escribano, J. P., Alt, S. and Bibb, M. J. (2016) 'Next generation sequencing of actinobacteria for the discovery of novel natural products', *Marine Drugs*. doi: 10.3390/md14040078.
Grant, J. R., Arantes, A. S. and Stothard, P. (2012) 'Comparing thousands of circular genomes using the CGView Comparison Tool', *BMC Genomics*. doi: 10.1186/1471-2164-13-202.
Greetham, D. (2014) 'Phenotype microarray technology and its application in industrial biotechnology', *Biotechnology letters*, 36. doi: 10.1007/s10529-014-1481-x.
Griffith, R. S. (1984) 'Vancomycin use--an historical review', *Journal of Antimicrobial Chemotherapy*. doi: 10.1093/jac/14.suppl_d.1.
Gu, H. *et al.* (1994) 'Deletion of a DNA polymerase β gene segment in T cells using cell type-specific gene targeting', *Science*. doi: 10.1126/science.8016642.
Gueldener, U. (2002) 'A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast', *Nucleic Acids Research*. doi: 10.1093/nar/30.6.e23.
Guelin, E., Rep, M. and Grivell, L. A. (1994) 'Yeast sequencing reports. Sequence of the AFG3 gene encoding a new member of the FtsH/Yme1/Tma subfamily of the AAA-protein family', *Yeast*. doi: 10.1002/yea.320101016.
van Gulik, W. M. and Heijnen, J. J. (1995) 'A metabolic network stoichiometry analysis of microbial growth and product formation', *Biotechnology and Bioengineering*. doi: 10.1002/bit.260480617.
Gurevich, A. *et al.* (2013) 'QUAST: Quality assessment tool for genome assemblies', *Bioinformatics*. doi: 10.1093/bioinformatics/btt086.
Gurung, N. *et al.* (2013) 'A broader view: Microbial enzymes and their relevance in industries, medicine, and beyond', *BioMed Research International*. doi: 10.1155/2013/329121.
Hale, T. W. (2019) 'Amoxicillin + Clavulanate', in *Hale's Medications & Mothers' Milk™ 2019*. doi: 10.1891/9780826150356.0047.
Harbor, C. S. (2010) 'YPD media', *Cold Spring Harb Protoc.* Available at: http://cshprotocols.cshlp.org/content/2010/9/pdb.rec12315.full.
Hata, T. *et al.* (1956) 'Mitomycin, a new antibiotic from Streptomyces. I.', *The Journal of antibiotics*.
Hata, T. and Sugawara, R. (1956) 'Mitomycin, a new antibiotic from Streptomyces. II. Description of the', *The Journal of antibiotics*. doi: 10.11554/antibioticsa.9.4_141.
Van Heel, A. J. *et al.* (2018) 'BAGEL4: A user-friendly web server to thoroughly mine RiPPs and bacteriocins', *Nucleic Acids Research*. doi: 10.1093/nar/gky383.
Heinemeyer, W. *et al.* (1991) 'Proteinase yscE, the yeast proteasome/multicatalytic-multifunctional proteinase: Mutants unravel its function in stress induced proteolysis and uncover its necessity for cell survival', *EMBO Journal*. doi: 10.1002/j.1460-2075.1991.tb07982.x.
Heitefuss, R. (2010) 'Secondary Metabolites in Soil Ecology', *Journal of Phytopathology*. doi: 10.1111/j.1439-0434.2009.01592.x.
Hentges, P. *et al.* (2005) 'Three novel antibiotic marker cassettes for gene disruption and marker switching in Schizosaccharomyces pombe', *Yeast*. doi: 10.1002/yea.1291.
Higgins, V. J. *et al.* (1999) 'Leu343Phe substitution in the Malx3 protein of Saccharomyces

cerevisiae increases the constitutivity and glucose insensitivity of MAL gene expression', *Current Genetics*. doi: 10.1007/s002940050444.
Hille, F. *et al.* (2018) 'The Biology of CRISPR-Cas: Backward and Forward', *Cell*. doi: 10.1016/j.cell.2017.11.032.
Himmel, M. E. *et al.* (2007) 'Biomass recalcitrance: Engineering plants and enzymes for biofuels production', *Science*. doi: 10.1126/science.1137016.
Hoff, K. J. *et al.* (2016) 'BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS', *Bioinformatics*. doi: 10.1093/bioinformatics/btv661.
Hoffman, C. S., Wood, V. and Fantes, P. A. (2015) 'An ancient yeast for young geneticists: A primer on the Schizosaccharomyces pombe model system', *Genetics*. doi: 10.1534/genetics.115.181503.
Hu, C. (2018) *Major pharmaceutical companies like Novartis dropping antibiotics projects and future tools against superbugs*, *Business Insider*.
Hyatt, D. *et al.* (2010) 'Prodigal: Prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*. doi: 10.1186/1471-2105-11-119.
Illumina (2014) 'Sequencing Methods Review (A review of publications featuring Illumina Technology)', *Illumina protocols*. doi: 10.1038/nmeth.2688.
Ishino, Y., Krupovic, M. and Forterre, P. (2018) 'History of CRISPR-Cas from encounter with a mysterious repeated sequence to genome editing technology', *Journal of Bacteriology*. doi: 10.1128/JB.00580-17.
Isken, O. and Maquat, L. E. (2009) 'Telomeric RNAs as a novel player in telomeric integrity', *F1000 Biology Reports*. doi: 10.3410/b1-90.
Jacq, C. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome IV', *Nature*.
Jain, C. *et al.* (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature Communications*. doi: 10.1038/s41467-018-07641-9.
Jasin, M. and Haber, J. E. (2016) 'The democratization of gene editing: Insights from site-specific cleavage and double-strand break repair', *DNA Repair*. doi: 10.1016/j.dnarep.2016.05.001.
JB, M. *et al.* (1997) 'Cloning, and molecular characterization of the GCV1 gene encoding the glycine cleavage T-protein from Saccharomyces cerevisiae.', *Gene*, 186(1), pp. 13–20.
JG, P. *et al.* (1988) 'Molecular genetics of serine and threonine catabolism in Saccharomyces cerevisiae.', *Genetics*, 119(3), pp. 527–534.
Ji, Z. *et al.* (2009) 'Isolation, structure elucidation and antibacterial activities of streptothricin acids', *Journal of Antibiotics*. doi: 10.1038/ja.2009.16.
Johnson, D. R. *et al.* (1994) 'Saccharomyces cerevisiae contains four fatty acid activation (FAA) genes: An assessment of their role in regulating protein N-myristoylation and cellular lipid metabolism', *Journal of Cell Biology*. doi: 10.1083/jcb.127.3.751.
Johnson, M. *et al.* (2008) 'NCBI BLAST: a better web interface.', *Nucleic acids research*. doi: 10.1093/nar/gkn201.
Johnston, M. *et al.* (1994) 'Complete nucleotide sequence of Saccharomyces cerevisiae chromosome VIII', *Science*. doi: 10.1126/science.8091229.
Johnston, M. (1996) 'Genome sequencing: The complete code for a eukaryotic cell', *Current Biology*. doi: 10.1016/S0960-9822(02)00526-2.
Jozala, A. F. *et al.* (2016) 'Biopharmaceuticals from microorganisms: from production to purification', *Brazilian Journal of Microbiology*. doi: 10.1016/j.bjm.2016.10.007.
Kamath, G. M. *et al.* (2017) 'HINGE: Long-read assembly achieves optimal repeat resolution',

Genome Research. doi: 10.1101/gr.216465.116.
Kamjam, M. *et al.* (2017) 'Deep sea actinomycetes and their secondary metabolites', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2017.00760.
Katz, L. and Baltz, R. H. (2016) 'Natural product discovery: past, present, and future', *Journal of Industrial Microbiology and Biotechnology*. doi: 10.1007/s10295-015-1723-5.
Kay, S. *et al.* (2007) 'A bacterial effector acts as a plant transcription factor and induces a cell size regulator', *Science*. doi: 10.1126/science.1144956.
Kellis, M., Birren, B. W. and Lander, E. S. (2004) 'Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae', *Nature*, 428 VN-(6983), pp. 617–624. doi: 10.1038/nature02424.
Kersten, H. (1975) 'Mechanism of Action of Mitomycins', in *Antineoplastic and Immunosuppressive Agents*. doi: 10.1007/978-3-642-65806-8_5.
Khaldi, N. *et al.* (2010) 'SMURF: Genomic mapping of fungal secondary metabolite clusters', *Fungal Genetics and Biology*. doi: 10.1016/j.fgb.2010.06.003.
Khan, M. I., Shin, J. H. and Kim, J. D. (2018) 'The promising future of microalgae: Current status, challenges, and optimization of a sustainable and renewable industry for biofuels, feed, and other products', *Microbial Cell Factories*. doi: 10.1186/s12934-018-0879-x.
Khokhlov, A. S. (1983) 'Advances in investigation of streptothricin antibiotics', *Antibiotiki*.
Kieser, T. *et al.* (1982) 'pIJ101, a multi-copy broad host-range Streptomyces plasmid: Functional analysis and development of DNA cloning vectors', *MGG Molecular & General Genetics*. doi: 10.1007/BF00330791.
Kim, E. J. *et al.* (2009) 'Expression of hepatitis B surface antigen S domain in recombinant Saccharomyces cerevisiae using GAL1 promoter', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2009.03.004.
Kinashi, H. (2011) 'Giant linear plasmids in Streptomyces: A treasure trove of antibiotic biosynthetic clusters', *Journal of Antibiotics*. doi: 10.1038/ja.2010.146.
Kirst, H. and Melis, A. (2014) 'The chloroplast signal recognition particle (CpSRP) pathway as a tool to minimize chlorophyll antenna size and maximize photosynthetic productivity', *Biotechnology Advances*. doi: 10.1016/j.biotechadv.2013.08.018.
Knothe, G. (2012) 'Fuel properties of highly polyunsaturated fatty acid methyl esters. Prediction of fuel properties of algal biodiesel', in *Energy and Fuels*. doi: 10.1021/ef300700v.
Knox, A. M., Du Preez, J. C. and Kilian, S. G. (2004) 'Starch fermentation characteristics of Saccharomyces cerevisiae strains transformed with amylase genes from Lipomyces kononenkoae and Saccharomycopsis fibuligera', *Enzyme and Microbial Technology*. doi: 10.1016/j.enzmictec.2003.12.010.
Koepff, J. *et al.* (2018) 'Germination and growth analysis of Streptomyces lividans at the single-cell level under varying medium compositions', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2018.02680.
Koizumi, T. (2014) 'Biofuels and food security', in *SpringerBriefs in Applied Sciences and Technology*. doi: 10.1007/978-3-319-05645-6_7.
Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature Biotechnology*. doi: 10.1038/s41587-019-0072-8.
Koren, S. *et al.* (2012) 'Hybrid error correction and de novo assembly of single-molecule sequencing reads', *Nature Biotechnology*. doi: 10.1038/nbt.2280.
Koren, S. *et al.* (2013) 'Reducing assembly complexity of microbial genomes with single-molecule sequencing', *Genome Biology*. doi: 10.1186/gb-2013-14-9-r101.
Koren, S. *et al.* (2017) 'Canu: Scalable and accurate long-read assembly via adaptive κ-mer

weighting and repeat separation', *Genome Research*. doi: 10.1101/gr.215087.116.
Kormanec, J. *et al.* (2019) 'Recent achievements in the generation of stable genome alterations/mutations in species of the genus Streptomyces', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-019-09901-0.
Kruckeberg, A. L. (1996) 'The hexose transporter family of Saccharomyces cerevisiae', *Archives of Microbiology*. doi: 10.1007/s002030050385.
Kubota, S. *et al.* (2004) 'Effect of ethanol on cell growth of budding yeast: Genes that are important for cell growth in the presence of ethanol', *Bioscience, Biotechnology and Biochemistry*. doi: 10.1271/bbb.68.968.
Kumar, K. R., Cowley, M. J. and Davis, R. L. (2019) 'Next-Generation Sequencing and Emerging Technologies', *Seminars in Thrombosis and Hemostasis*. doi: 10.1055/s-0039-1688446.
Kvamme, E. (1958) 'Glycolysis and Respiration in Ehrlich Ascites Tumour Cells. I. Phosphate Metabolism in Relation to Glycolysis and the Crabtree Effect.', *Acta Physiologica Scandinavica*. doi: 10.1111/j.1748-1716.1958.tb01553.x.
Kwak, S. and Jin, Y. S. (2017) 'Production of fuels and chemicals from xylose by engineered Saccharomyces cerevisiae: A review and perspective', *Microbial Cell Factories*. doi: 10.1186/s12934-017-0694-9.
Lagunas, R. (1976) 'Energy metabolism of Saccharomyces cerevisiae discrepancy between ATP balance and known metabolic functions', *BBA - Bioenergetics*. doi: 10.1016/0005-2728(76)90049-9.
Lanfear, R. *et al.* (2019) 'MinIONQC: Fast and simple quality control for MinION sequencing data', *Bioinformatics*. doi: 10.1093/bioinformatics/bty654.
Langmead, B. (2014) 'De Bruijn Graph assembly', *Journal of Software*.
Langmead, B. and Salzberg, S. (2013) 'Bowtie2', *Nature methods*. doi: 10.1038/nmeth.1923.Fast.
Langreth, R. (2019) 'Antibiotics Aren't Profitable Enough for Big Pharma to Make More', *Bloomberg Businessweek*.
Lardenois, A. *et al.* (2011) 'Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1016459108.
Larkin, M. A. *et al.* (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*. doi: 10.1093/bioinformatics/btm404.
Larroy, C. *et al.* (2002) 'Characterization of the Saccharomyces cerevisiae YMR318c (ADH6) gene product as a broad specificity NADPH-dependent alcohol dehydrogenase: Relevance in aldehyde reduction', *Biochemical Journal*. doi: 10.1042/0264-6021:3610163.
Lees, J. A. *et al.* (2016) 'Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes', *Nature Communications*. doi: 10.1038/ncomms12797.
Lees, J. A. *et al.* (2018) 'pyseer: A comprehensive tool for microbial pangenome-wide association studies', *Bioinformatics*. doi: 10.1093/bioinformatics/bty539.
Li, S. *et al.* (2013) 'Eukaryotic resistance to fluoride toxicity mediated by a widespread family of fluoride export proteins', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1310439110.
Li, T. *et al.* (2011) 'Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes', *Nucleic Acids Research*. doi: 10.1093/nar/gkr188.
Liao, Y., Smyth, G. K. and Shi, W. (2013) 'The Subread aligner: Fast, accurate and scalable read

mapping by seed-and-vote', *Nucleic Acids Research*. doi: 10.1093/nar/gkt214.
Lim, S. K. *et al.* (2009) 'Iso-migrastatin, migrastatin, and dorrigocin production in streptomyces platensis NRRL 18993 is governed by a single biosynthetic machinery featuring an acyltransferase-less type I polyketide synthase', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M109.046805.
de Lima Procópio, R. E. *et al.* (2012) 'Antibiotics produced by Streptomyces', *Brazilian Journal of Infectious Diseases*. doi: 10.1016/j.bjid.2012.08.014.
Limayem, A. and Ricke, S. C. (2012) 'Lignocellulosic biomass for bioethanol production: Current perspectives, potential issues and future prospects', *Progress in Energy and Combustion Science*. doi: 10.1016/j.pecs.2012.03.002.
Lindegren, C. C. (2011) *The yeast cell, its genetics and cytology.*, *The yeast cell, its genetics and cytology.* doi: 10.5962/bhl.title.7236.
Lippert, C. *et al.* (2011) 'FaST linear mixed models for genome-wide association studies', *Nature Methods*. doi: 10.1038/nmeth.1681.
Lischer, H. E. L. and Shimizu, K. K. (2017) 'Reference-guided de novo assembly approach improves genome reconstruction for related species', *BMC Bioinformatics*. doi: 10.1186/s12859-017-1911-6.
Little, A. E. F. *et al.* (2008) 'Rules of Engagement: Interspecies Interactions that Regulate Microbial Communities', *Annual Review of Microbiology*. doi: 10.1146/annurev.micro.030608.101423.
Lowe, T. M. and Eddy, S. R. (1996) 'TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence', *Nucleic Acids Research*. doi: 10.1093/nar/25.5.0955.
Lundquist, T. J. *et al.* (2010) 'A ! Realistic ! Technology ! and ! Engineering ! Assessment ! of ! Algae ! Biofuel ! Production !', *Energy*.
Luo, Z. and van Vuuren, H. J. J. (2009) 'Functional analyses of PAU genes in Saccharomyces cerevisiae', *Microbiology*. doi: 10.1099/mic.0.030726-0.
Ma, M. and Liu, Z. L. (2010) 'Mechanisms of ethanol tolerance in saccharomyces cerevisiae', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-010-2594-3.
Mack, A. *et al.* (1992) 'Stoichiometry of the cre recombinase bound to the lox recombining site', *Nucleic Acids Research*. doi: 10.1093/nar/20.17.4451.
Mackie, A. M. *et al.* (2014) 'Biolog phenotype microarrays for phenotypic characterization of microbial cells', *Methods in Molecular Biology*. doi: 10.1007/978-1-62703-712-9_10.
Mäkinen, V., Salmela, L. and Ylinen, J. (2012) 'Normalized N50 assembly metric using gap-restricted co-linear chaining', *BMC Bioinformatics*. doi: 10.1186/1471-2105-13-255.
von der Malsburg, K. *et al.* (2011) 'Dual Role of Mitofilin in Mitochondrial Membrane Organization and Protein Biogenesis', *Developmental Cell*. doi: 10.1016/j.devcel.2011.08.026.
Market and Market (2017) *Bioethanol Market by Feedstock (Starch-Based, Sugar-Based, Cellulose-Based), End-Use Industry (Transportation, Pharmaceuticals, Cosmetics, Alcoholic Beverages), Blend (E5, E10, E15 to E70, E75 to E85), and Region - Global Forecast to 2022*.
Martin, M. A. (2010) 'First generation biofuels compete', *New Biotechnology*. doi: 10.1016/j.nbt.2010.06.010.
Maskey, R. P. *et al.* (2004) 'Anti-cancer and antibacterial trioxacarcins with high anti-malaria activity from a marine streptomycete and their absolute stereochemistry', *Journal of Antibiotics*. doi: 10.7164/antibiotics.57.771.
Mattanovich, D., Sauer, M. and Gasser, B. (2014) 'Yeast biotechnology: Teaching the old dog new tricks', *Microbial Cell Factories*. doi: 10.1186/1475-2859-13-34.

Maxam, A. M. and Gilbert, W. (1977) 'A new method for sequencing DNA', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.74.2.560.
McNaughton, A. L. *et al.* (2019) 'Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV)', *Scientific Reports*. doi: 10.1038/s41598-019-43524-9.
Medema, M. H. *et al.* (2011) 'AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences', *Nucleic Acids Research*. doi: 10.1093/nar/gkr466.
Mekonnen, M. M. *et al.* (2018) 'Water, Energy, and Carbon Footprints of Bioethanol from the U.S. and Brazil', *Environmental Science and Technology*. doi: 10.1021/acs.est.8b03359.
Messer, W. *et al.* (2001) 'Bacterial replication initiator DnaA. Rules for DnaA binding and roles of DnaA in origin unwinding and helicase loading', *Biochimie*. doi: 10.1016/S0300-9084(00)01216-5.
Messer, W. (2002) 'The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication', *FEMS Microbiology Reviews*. doi: 10.1016/S0168-6445(02)00127-4.
Miller, J. R., Koren, S. and Sutton, G. (2010) 'Assembly algorithms for next-generation sequencing data', *Genomics*. doi: 10.1016/j.ygeno.2010.03.001.
Milne, I. *et al.* (2016) 'Tablet: Visualizing next-generation sequence assemblies and mappings', in *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-3167-5_14.
Min, B., Grigoriev, I. V. and Choi, I. G. (2017) 'FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation', *Bioinformatics*. doi: 10.1093/bioinformatics/btx353.
Miné-Hattab, J. and Rothstein, R. (2013) 'Gene targeting and homologous recombination in saccharomyces cerevisiae', *Topics in Current Genetics*. doi: 10.1007/978-94-007-4531-5_3.
Moir, D. T. *et al.* (1999) 'Genomics and antimicrobial drug discovery', *Antimicrobial Agents and Chemotherapy*. doi: 10.1128/aac.43.3.439.
Moon, S. K., Kim, S. W. and Choi, G. W. (2012) 'Simultaneous saccharification and continuous fermentation of sludge-containing mash for bioethanol production by Saccharomyces cerevisiae CHFY0321', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2011.06.009.
Morisse, P., Lecroq, T. and Lefebvre, A. (2020) 'Long-read error correction: a survey and qualitative comparison', *bioRxiv*. doi: 10.1101/2020.03.06.977975.
Mortimer, R. K. and Johnston, J. R. (1986) 'Genealogy of principal strains of the yeast genetic stock center', *Genetics*.
Mu, R. *et al.* (2020) 'Application of enzymes in regioselective and stereoselective organic reactions', *Catalysts*. doi: 10.3390/catal10080832.
Myers, E. W. (2005) 'The fragment assembly string graph', *Bioinformatics*. doi: 10.1093/bioinformatics/bti1114.
Nagalakshmi, U. *et al.* (2008) 'The transcriptional landscape of the yeast genome defined by RNA sequencing', *Science*. doi: 10.1126/science.1158441.
Naik, S. N. *et al.* (2010) 'Production of first and second generation biofuels: A comprehensive review', *Renewable and Sustainable Energy Reviews*. doi: 10.1016/j.rser.2009.10.003.
Nakamura, K. *et al.* (2011) 'Sequence-specific error profile of Illumina sequencers', *Nucleic Acids Research*. doi: 10.1093/nar/gkr344.
Narala, R. R. *et al.* (2016) 'Comparison of microalgae cultivation in photobioreactor, open raceway pond, and a two-stage hybrid system', *Frontiers in Energy Research*. doi: 10.3389/fenrg.2016.00029.
Nazhad, M. M. *et al.* (1995) 'Structural constraints affecting the initial enzymatic hydrolysis of

recycled paper', *Enzyme and Microbial Technology*. doi: 10.1016/0141-0229(94)00057-X.
NCBI Resource Coordinators (2017) 'Database Resources of the National Center for Biotechnology Information.', *Nucleic acids research*. doi: 10.1093/nar/gkw1071.
Neiman, A. M. (2005) 'Ascospore Formation in the Yeast Saccharomyces cerevisiae', *Microbiology and Molecular Biology Reviews*. doi: 10.1128/mmbr.69.4.565-584.2005.
Nilsson, A. and Nielsen, J. (2016) 'Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase.', *Scientific reports*, 6, p. 22264. doi: 10.1038/srep22264.
Nishida, H. (2006) 'Detection and characterization of fungal-specific proteins in Saccharomyces cerevisiae', *Bioscience, Biotechnology and Biochemistry*. doi: 10.1271/bbb.60251.
Oh, H. N. *et al.* (2019) 'Antarctic tundra soil metagenome as useful natural resources of cold-active lignocelluolytic enzymes', *Journal of Microbiology*. doi: 10.1007/s12275-019-9217-1.
Ohkuni, K. *et al.* (2006) 'Suppressor analysis of the mpt5/htr1/uth4/puf5 deletion in Saccharomyces cerevisiae', *Molecular Genetics and Genomics*. doi: 10.1007/s00438-005-0064-x.
Okada, B. K. and Seyedsayamdost, M. R. (2017) 'Antibiotic dialogues: Induction of silent biosynthetic gene clusters by exogenous small molecules', *FEMS Microbiology Reviews*. doi: 10.1093/femsre/fuw035.
Okoro, C. K. *et al.* (2009) 'Diversity of culturable actinomycetes in hyper-arid soils of the Atacama Desert, Chile', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*. doi: 10.1007/s10482-008-9295-2.
Ondov, B. D. *et al.* (2016) 'Mash: Fast genome and metagenome distance estimation using MinHash', *Genome Biology*. doi: 10.1186/s13059-016-0997-x.
Orr-Weaver, T. L., Szostak, J. W. and Rothstein, R. J. (1981) 'Yeast transformation: A model system for the study of recombination', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.78.10.6354.
Overbeek, R. *et al.* (2014) 'The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)', *Nucleic Acids Research*. doi: 10.1093/nar/gkt1226.
Palkova, Z. (2004) 'Multicellular microorganisms: Laboratory versus nature', *EMBO Reports*. doi: 10.1038/sj.embor.7400145.
Pandey, A. K. *et al.* (2019) 'Evaluation of divergent yeast genera for fermentation-associated stresses and identification of a robust sugarcane distillery waste isolate Saccharomyces cerevisiae NGY10 for lignocellulosic ethanol production in SHF and SSF', *Biotechnology for Biofuels*. doi: 10.1186/s13068-019-1379-x.
Pandza, K. *et al.* (1997) 'Physical mapping shows that the unstable oxytetracycline gene cluster of Streptomyces rimosus lies close to one end of the linear chromosome', *Microbiology*. doi: 10.1099/00221287-143-5-1493.
Paradkar, A. *et al.* (2003) 'Streptomyces genetics: A genomic perspective', *Critical Reviews in Biotechnology*. doi: 10.1080/713609296.
Parenteau, J. *et al.* (2008) 'Deletion of many yeast introns reveals a minority of genes that require splicing for function', *Molecular Biology of the Cell*. doi: 10.1091/mbc.E07-12-1254.
Paul Abishek, M., Patel, J. and Prem Rajan, A. (2014) 'Algae Oil: A Sustainable Renewable Fuel of Future', *Biotechnology Research International*. doi: 10.1155/2014/272814.
Pavlopoulos, G. A. *et al.* (2015) 'Visualizing genome and systems biology: Technologies, tools, implementation techniques and trends, past, present and future', *GigaScience*. doi: 10.1186/s13742-015-0077-2.
Pera, L. M. *et al.* (2015) 'Biocatalysis', in *Industrial Biorefineries and White Biotechnology*. doi:

10.1016/B978-0-444-63453-5.00012-4.
Peralta-Yahya, P. P. and Keasling, J. D. (2010) 'Advanced biofuel production in microbes', *Biotechnology Journal*. doi: 10.1002/biot.200900220.
Pfeiffer, T., Schuster, S. and Bonhoeffer, S. (2001) 'Cooperation and competition in the evolution of ATP-producing pathways.', *Science (New York, N.Y.)*, 292(April), pp. 504–507. doi: 10.1126/science.1058079.
Philippsen, P. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome XIV and its evolutionary implications.', *Nature*. doi: 10.1038/387s093.
Pina, C. *et al.* (2004) 'Ethanol tolerance of five non-Saccharomyces wine yeasts in comparison with a strain of Saccharomyces cerevisiae - Influence of different culture conditions', *Food Microbiology*. doi: 10.1016/j.fm.2003.10.009.
Piškur, J. *et al.* (2006) 'How did Saccharomyces evolve to become a good brewer?', *Trends in Genetics*. doi: 10.1016/j.tig.2006.02.002.
Plessis, A. *et al.* (1992) 'Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus', *Genetics*.
Postma, E. *et al.* (1989) 'Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae.', *Applied and Environmental Microbiology*, 55(2), pp. 468–477.
Power, R. A., Parkhill, J. and De Oliveira, T. (2016) 'Microbial genome-wide association studies: lessons from human GWAS', *Nature Reviews Genetics*. doi: 10.1038/nrg.2016.132.
Prokop, J. W. *et al.* (2018) 'Genome sequencing in the clinic: The past, present, and future of genomic medicine', *Physiological Genomics*. doi: 10.1152/physiolgenomics.00046.2018.
Proux-Wéra, E. *et al.* (2012) 'A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach', *BMC Bioinformatics*. doi: 10.1186/1471-2105-13-237.
Puetz, J. and Wurm, F. M. (2019) 'Recombinant Proteins for Industrial versus Pharmaceutical Purposes: A Review of Process and Pricing', *Processes*. doi: 10.3390/pr7080476.
Purcell, S. *et al.* (2007) 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *American Journal of Human Genetics*. doi: 10.1086/519795.
RAKE, G. and DONOVICK, R. (1946) 'Studies on the nutritional requirements of Actinomyces griseus for the formation of streptomycin.', *Journal of bacteriology*.
Ramos, J., Szkutnicka, K. and VP, C. (1989) 'Characteristics of galactose transport in Saccharomyces cerevisiae cells and reconstituted lipid vesicles.', *Journal of bacteriology*, 171(6), pp. 3539–3544.
Rang, F. J., Kloosterman, W. P. and de Ridder, J. (2018) 'From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy', *Genome Biology*. doi: 10.1186/s13059-018-1462-9.
Ray, R. C. (2018) *Bioethanol Production from Food Crops: Sustainable Sources, Interventions, and Challenges.* San Diego: Elsevier Science & Technology.
Ribeiro, O. *et al.* (2007) 'Application of the Cre-loxP system for multiple gene disruption in the yeast Kluyveromyces marxianus', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2007.05.027.
Richard, P., MH, T. and Penttilä, M. (1999) 'Evidence that the gene YLR070c of Saccharomyces cerevisiae encodes a xylitol dehydrogenase.', *FEBS letters*, 457(1), pp. 135–138.
Riles, L. and Fay, J. C. (2019) 'Genetic basis of variation in heat and ethanol tolerance in saccharomyces cerevisiae', *G3: Genes, Genomes, Genetics*. doi: 10.1534/g3.118.200566.
Robinson, J. T. *et al.* (2011) 'Integrative genomics viewer', *Nature Biotechnology*. doi:

10.1038/nbt.1754.
Robinson, J. T. *et al.* (2017) 'Variant review with the integrative genomics viewer', *Cancer Research*. doi: 10.1158/0008-5472.CAN-17-0337.
Römer, P. *et al.* (2007) 'Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene', *Science*. doi: 10.1126/science.1144958.
Rothstein, R. J. (1983) 'One-Step Gene Disruption in Yeast', *Methods in Enzymology*. doi: 10.1016/0076-6879(83)01015-0.
Ruangsomboon, S. *et al.* (2020) 'Outdoor open pond batch production of green microalga Botryococcus braunii for high hydrocarbon production: enhanced production with salinity', *Scientific Reports*. doi: 10.1038/s41598-020-59645-5.
Russnak, R., Konczal, D. and McIntire, S. L. (2001) 'A Family of Yeast Proteins Mediating Bidirectional Vacuolar Amino Acid Transport', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M008028200.
Rutherford, K. *et al.* (2000) 'Artemis: Sequence visualization and annotation', *Bioinformatics*. doi: 10.1093/bioinformatics/16.10.944.
Sa-uth, C. *et al.* (2018) 'Modification of Medium Composition for Enhancing the Production of Antifungal Activity from *Xenorhabdus stockiae* PB09 by Using Response Surface Methodology', *International Journal of Microbiology*. Edited by G. Comi. Hindawi, 2018, p. 3965851. doi: 10.1155/2018/3965851.
Salzberg, S. L. *et al.* (1998) 'Microbial gene identification using interpolated Markov models', *Nucleic Acids Research*. doi: 10.1093/nar/26.2.544.
San, J. E. *et al.* (2020) 'Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2019.03119.
Sanger, F. and Coulson, A. R. (1975) 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', *Journal of Molecular Biology*. doi: 10.1016/0022-2836(75)90213-2.
Sanjana, N. E. *et al.* (2012) 'A transcription activator-like effector toolbox for genome engineering', *Nature Protocols*. doi: 10.1038/nprot.2011.431.
Santos-Aberturas, J. *et al.* (2019) 'Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool', *Nucleic Acids Research*. doi: 10.1093/nar/gkz192.
Saudagar, P. S., Survase, S. A. and Singhal, R. S. (2008) 'Clavulanic acid: A review', *Biotechnology Advances*. doi: 10.1016/j.biotechadv.2008.03.002.
Sayers, E. W. *et al.* (2009) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*. doi: 10.1093/nar/gkn741.
Schenk, P. M. *et al.* (2008) 'Second Generation Biofuels: High-Efficiency Microalgae for Biodiesel Production', *BioEnergy Research*. doi: 10.1007/s12155-008-9008-8.
Schwann, T. (1836) 'Ueber das Wesen des Verdauungsprocesses', *Annalen der Physik*. doi: 10.1002/andp.18361140612.
Schwartz, K., Richards, K. and Botstein, D. (1997) 'BIM1 encodes a microtubule-binding protein in yeast', *Molecular Biology of the Cell*. doi: 10.1091/mbc.8.12.2677.
Schweikert, G. *et al.* (2009) 'mGene: Accurate SVM-based gene finding with an application to nematode genomes', *Genome Research*. doi: 10.1101/gr.090597.108.
Sedlazeck, F. J. *et al.* (2018) 'Accurate detection of complex structural variations using single-molecule sequencing', *Nature Methods*. doi: 10.1038/s41592-018-0001-7.
Seemann, T. (2014) 'Prokka: Rapid prokaryotic genome annotation', *Bioinformatics*. doi:

10.1093/bioinformatics/btu153.
Seghetta, M. *et al.* (2016) 'Life cycle assessment of macroalgal biorefinery for the production of ethanol, proteins and fertilizers – A step towards a regenerative bioeconomy', *Journal of Cleaner Production*. doi: 10.1016/j.jclepro.2016.07.195.
Seipke, R. F. *et al.* (2012) 'Isolating antifungals from fungus-growing ant symbionts using a genome-guided chemistry approach', in *Methods in Enzymology*. doi: 10.1016/B978-0-12-404634-4.00003-6.
Serretta, V. *et al.* (2016) 'Mitomycin C from birth to adulthood', *Urologia*. doi: 10.5301/uro.5000195.
Sheldon, P. J. *et al.* (1999) 'Mitomycin resistance in Streptomyces lavendulae includes a novel drug- binding-protein-dependent export system', *Journal of Bacteriology*. doi: 10.1128/jb.181.8.2507-2512.1999.
Shigechi, H. *et al.* (2004) 'Direct production of ethanol from raw corn starch via fermentation by use of a novel surface-engineered yeast strain codisplaying glucoamylase and α-amylase', *Applied and Environmental Microbiology*. doi: 10.1128/AEM.70.8.5037-5040.2004.
Shiratori, A. *et al.* (1999) 'Systematic identification, classification, and characterization of the open reading frames which encode novel helicase-related proteins in Saccharomyces cerevisiae by gene disruption and Northern analysis', *Yeast*. doi: 10.1002/(SICI)1097-0061(199902)15:3<219::AID-YEA349>3.0.CO;2-3.
Show, P. L. *et al.* (2017) 'A holistic approach to managing microalgae for biofuel applications', *International Journal of Molecular Sciences*. doi: 10.3390/ijms18010215.
Shulga, N. *et al.* (1999) 'A nuclear export signal prevents Saccharomyces cerevisiae Hsp70 Ssb1p from stimulating nuclear localization signal-directed nuclear transport', *Journal of Biological Chemistry*. doi: 10.1074/jbc.274.23.16501.
Sievers, F. and Higgins, D. G. (2014) 'Clustal Omega', *Current Protocols in Bioinformatics*. doi: 10.1002/0471250953.bi0313s48.
Silver, L. L. (2011) 'Challenges of antibacterial discovery', *Clinical Microbiology Reviews*. doi: 10.1128/CMR.00030-10.
Simão, F. A. *et al.* (2015) 'BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*. doi: 10.1093/bioinformatics/btv351.
Sims, R. E. H. *et al.* (2010) 'An overview of second generation biofuel technologies', *Bioresource Technology*. doi: 10.1016/j.biortech.2009.11.046.
Singh, A., Nigam, P. S. and Murphy, J. D. (2011) 'Renewable fuels from algae: An answer to debatable land based fuels', *Bioresource Technology*. doi: 10.1016/j.biortech.2010.06.032.
Singh, J. and Gu, S. (2010) 'Commercialization potential of microalgae for biofuels production', *Renewable and Sustainable Energy Reviews*. doi: 10.1016/j.rser.2010.06.014.
Skinnider, M. A. *et al.* (2015) 'Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM)', *Nucleic Acids Research*. doi: 10.1093/nar/gkv1012.
Smalla, K. *et al.* (1998) 'Analysis of BIOLOG GN substrate utilization patterns by microbial communities', *Applied and Environmental Microbiology*. doi: 10.1128/aem.64.4.1220-1225.1998.
SMITH, D. W. and AMES, B. N. (1965) 'Phosphoribosyladenosine Monophosphate, an Intermediate in Histidine Biosynthesis.', *The Journal of biological chemistry*. Â© 1965 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 240(7), pp. 3056–3063. doi: 10.1016/s0021-9258(18)97286-8.
Snoek, T., Verstrepen, K. J. and Voordeckers, K. (2016) 'How do yeast cells become tolerant to high ethanol concentrations?', *Current Genetics*. doi: 10.1007/s00294-015-0561-3.

Soares, E. V. (2011) 'Flocculation in Saccharomyces cerevisiae: A review', *Journal of Applied Microbiology*, pp. 1–18. doi: 10.1111/j.1365-2672.2010.04897.x.
Son, M. and Taylor, R. (2012) 'Growth and Maintenance of Escherichia coli Laboratory Strains', *Current protocols in microbiology*, Chapter 5, p. Unit5A.4. doi: 10.1002/9780471729259.mc05a04s27.
Song, H. *et al.* (2017) 'High purity recombinant human growth hormone (rhGH) expression in Escherichia coli under phoA promoter', *Bioengineered*. doi: 10.1080/21655979.2016.1212137.
Spagnuolo, M., Yaguchi, A. and Blenner, M. (2019) 'Oleaginous yeast for biofuel and oleochemical production', *Current Opinion in Biotechnology*. doi: 10.1016/j.copbio.2019.02.011.
Spode, I. *et al.* (2002) 'ATF/CREB sites present in sub-telomeric regions of Saccharomyces cerevisiae chromosomes are part of promoters and act as UAS/URS of highly conserved COS genes', *Journal of Molecular Biology*. doi: 10.1016/S0022-2836(02)00322-4.
Stanke, M. *et al.* (2006) 'AUGUSTUS: A b initio prediction of alternative transcripts', *Nucleic Acids Research*. doi: 10.1093/nar/gkl200.
Stolz, L. E. *et al.* (1998) 'Identification and characterization of an essential family of inositol polyphosphate 5-phosphatases (INP51, INP52 and INP53 gene products) in the yeast Saccharomyces cerevisiae', *Genetics*.
Storici, F. *et al.* (2003) 'Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.2036296100.
Stovicek, V., Holkenbrink, C. and Borodina, I. (2017) 'CRISPR/Cas system for yeast genome engineering: advances and applications', *FEMS yeast research*. doi: 10.1093/femsyr/fox030.
Subramani, R. and Aalbersberg, W. (2013) 'Culturable rare Actinomycetes: Diversity, isolation and marine natural product discovery', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-013-5229-7.
Subramani, R. and Sipkema, D. (2019) 'Marine rare actinomycetes: A promising source of structurally diverse and unique novel natural products', *Marine Drugs*. doi: 10.3390/md17050249.
Sudha Rani, K. *et al.* (1996) 'Improved ethanol tolerance and production in strains of Clostridium thermocellum', *World Journal of Microbiology and Biotechnology*. doi: 10.1007/bf00327802.
Tam, V. *et al.* (2019) 'Benefits and limitations of genome-wide association studies', *Nature Reviews Genetics*. doi: 10.1038/s41576-019-0127-1.
Tapiero, J. (2008) 'Clavulanic acid', in *xPharm: The Comprehensive Pharmacology Reference*. doi: 10.1016/B978-008055232-3.63201-5.
Teixeira, M. T. (2013) 'Saccharomyces cerevisiae as a model to study replicative senescence triggered by telomere shortening', *Frontiers in Oncology*. doi: 10.3389/fonc.2013.00101.
Teng, H. *et al.* (2018) 'Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning', *GigaScience*. doi: 10.1093/gigascience/giy037.
Teresa Avelar, A. *et al.* (2013) 'Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy', *Nature Communications*. doi: 10.1038/ncomms3235.
Teste, M. A., Marie François, J. and Parrou, J. L. (2010) 'Characterization of a new multigene family encoding isomaltases in the yeast Saccharomyces cerevisiae, the IMA family', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M110.145946.
Tettelin, H *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome

VII.', *Nature*.
Tettelin, H. *et al.* (1997) 'The nucleotide sequence of Saccharomyces cerevisiae chromosome XV', *Nature*. doi: 10.1038/387s098.
Thaler, D. (1988) 'DNA Double-Chain Breaks In Recombination Of Phage Lambda And Of Yeast', *Annual Review of Genetics*. doi: 10.1146/annurev.genet.22.1.169.
The Free Library (2006) *World's oldest beer receipt?*, *Business Journals, Inc.* Available at: https://www.thefreelibrary.com/World%27s+oldest+beer+receipt%3F-a0141346971 (Accessed: 13 March 2021).
Thomson, J. M. *et al.* (2005) 'Resurrecting ancestral alcohol dehydrogenases from yeast', *Nature Genetics*. doi: 10.1038/ng1553.
Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration', *Briefings in Bioinformatics*. doi: 10.1093/bib/bbs017.
Tilak, M. K. *et al.* (2018) 'Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA', *Genome Biology and Evolution*. doi: 10.1093/gbe/evy022.
Timmis, K. *et al.* (2016) 'Microbial Biotechnology-2020', *Microbial biotechnology*. 2016/08/11. John Wiley and Sons Inc., 9(5), p. 529. doi: 10.1111/1751-7915.12403.
Tørresen, O. K. *et al.* (2019) 'Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases', *Nucleic acids research*. doi: 10.1093/nar/gkz841.
Tsunoda, T., Tanaka, T. and Nakamura, Y. (2019) *Genome-wide association studies*, *Genome-Wide Association Studies*. doi: 10.1007/978-981-13-8177-5.
Urban-Grimal, D. and Labbe-Bois, R. (1981) 'Genetic and biochemical characterization of mutants of Saccharomyces cerevisiae blocked in six different steps of heme biosynthesis', *MGG Molecular & General Genetics*. doi: 10.1007/BF00270144.
Vardi, N. (2005) 'Moore's law', *Forbes*. doi: 10.1201/noe0849396397.ch213.
Vaz Jauri, P. *et al.* (2013) 'Subinhibitory antibiotic concentrations mediate nutrient use and competition among soil streptomyces', *PLoS ONE*. doi: 10.1371/journal.pone.0081064.
Velmurugan, R. and Incharoensakdi, A. (2020) 'Heterologous Expression of Ethanol Synthesis Pathway in Glycogen Deficient Synechococcus elongatus PCC 7942 Resulted in Enhanced Production of Ethanol and Exopolysaccharides', *Frontiers in Plant Science*. doi: 10.3389/fpls.2020.00074.
Vembar, S. S. *et al.* (2016) 'Complete telomere-to-telomere de novo assembly of the Plasmodiumfalciparumgenome through long-read (>11 kb), single molecule, real-time sequencing', *DNA Research*. doi: 10.1093/dnares/dsw022.
Venkata Mohan, S. *et al.* (2016) 'A Circular Bioeconomy with Biobased Products from CO2 Sequestration', *Trends in Biotechnology*. doi: 10.1016/j.tibtech.2016.02.012.
Venter, U. and Hörz, W. (1989) 'The acid phosphatase genes PHO10 and PHO11 in S.cerevisiae are located at the telomeres of chromosomes VIII and I', *Nucleic Acids Research*. doi: 10.1093/nar/17.4.1353.
Vera, J. M. and Dowell, R. D. (2016) 'Survey of cryptic unstable transcripts in yeast', *BMC Genomics*. doi: 10.1186/s12864-016-2622-5.
Verduyn, C. *et al.* (1991) 'A theoretical evaluation of growth yields of yeasts', *Antonie van Leeuwenhoek*. doi: 10.1007/BF00582119.
Verweij, J. and Pinedo, H. M. (1990) 'Mitomycin C: mechanism of action, usefulness and limitations.', *Anti-cancer drugs*. doi: 10.1097/00001813-199010000-00002.
Větrovský, T. and Baldrian, P. (2013) 'The Variability of the 16S rRNA Gene in Bacterial

Genomes and Its Consequences for Bacterial Community Analyses', *PLoS ONE*. doi: 10.1371/journal.pone.0057923.
Vig, B. K. and Lewis, R. (1978) 'Genetic toxicology of bleomycin', *Mutation Research/Reviews in Genetic Toxicology*. doi: 10.1016/0165-1110(78)90019-2.
Vijayaraj Reddy, M. and Randerath, K. (1987) '32P-analysis of DNA adducts in somatic and reproductive tissues of rats treated with the anticancer antibiotic, mitomycin C', *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*. doi: 10.1016/0027-5107(87)90043-1.
Vijayraghavan, U., Company, M. and Abelson, J. (1989) 'Isolation and characterization of pre-mRNA splicing mutants of Saccharomyces cerevisiae.', *Genes & development*. doi: 10.1101/gad.3.8.1206.
De Villiers, T. (2008) 'Fungal enzymes and microbial systems for industrial processing', *Stellenbosch University*.
WAKSMAN, S. A., HARRIS, D. and LECHEVALIER, M. (1951) 'Studies on Streptomyces lavendulae.', *Journal of bacteriology*. doi: 10.1128/JB.62.2.149-161.1951.
Waksman, S. A. and Lechevalier, H. A. (1949) 'Neomycin, a new antibiotic active against streptomycin-resistant bacteria, including tuberculosis organisms', *Science*. doi: 10.1126/science.109.2830.305.
Waksman, S. A. and Woodruff, H. B. (1942) 'Streptothricin, a New Selective Bacteriostatic and Bactericidal Agent, Particularly Active Against Gram-Negative Bacteria', *Proceedings of the Society for Experimental Biology and Medicine*. doi: 10.3181/00379727-49-13515.
Wallace-Salinas, V. *et al.* (2015) 'Cell periphery-related proteins as major genomic targets behind the adaptive evolution of an industrial Saccharomyces cerevisiae strain to combined heat and hydrolysate stress', *BMC Genomics*. doi: 10.1186/s12864-015-1737-4.
Walter, J. M. *et al.* (2019) 'Method for Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas9', in *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-9736-7_3.
Wang, H., La Russa, M. and Qi, L. S. (2016) 'CRISPR/Cas9 in Genome Editing and Beyond', *Annual Review of Biochemistry*. doi: 10.1146/annurev-biochem-060815-014607.
Weber, T. *et al.* (2009) 'CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters', *Journal of Biotechnology*. doi: 10.1016/j.jbiotec.2009.01.007.
Wellinger, R. J., Wolf, A. J. and Zakian, V. A. (1993) 'Saccharomyces telomeres acquire single-strand TG1-3 tails late in S phase', *Cell*. doi: 10.1016/0092-8674(93)90049-V.
Wellinger, R. J. and Zakian, V. A. (2012) 'Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: Beginning to end', *Genetics*. doi: 10.1534/genetics.111.137851.
Wetterstrand KA (no date) *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Available at: www.genome.gov/sequencingcostsdata. (Accessed: 13 June 2020).
Wick, R. R. *et al.* (2017) 'Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads', *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1005595.
Wick, R. R. and Holt, K. E. (2019) 'Benchmarking of long-read assemblers for prokaryote whole genome sequencing', *F1000Research*. doi: 10.12688/f1000research.21782.1.
Wick, R. R., Judd, L. M. and Holt, K. E. (2019) 'Performance of neural network basecalling tools for Oxford Nanopore sequencing', *Genome Biology*. doi: 10.1186/s13059-019-1727-y.
Wigge, P. A. *et al.* (1998) 'Analysis of the Saccharomyces spindle pole by matrix-assisted laser

desorption/ionization (MALDI) mass spectrometry', *Journal of Cell Biology*. doi: 10.1083/jcb.141.4.967.
Wijffels, R. H. and Barbosa, M. J. (2010) 'An outlook on microalgal biofuels', *Science*. doi: 10.1126/science.1189003.
Wilson, D. B. *et al.* (2020) *Industrial Microbiology*. Wiley. Available at: https://books.google.co.uk/books?id=7w6_DwAAQBAJ.
Wolfe, K. H. and Shields, D. C. (1997) 'Molecular evidence for an ancient duplication of the entire yeast genome', *Nature*. doi: 10.1038/42711.
Wu, J., Delneri, D. and O'Keefe, R. T. (2012) 'Non-coding RNAs in Saccharomyces cerevisiae: What is the function?', in *Biochemical Society Transactions*. doi: 10.1042/BST20120042.
Xiao, C. Le *et al.* (2017) 'MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads', *Nature Methods*. doi: 10.1038/nmeth.4432.
Yamada, M. *et al.* (1998) 'Y'-help 1, a DNA helicase encoded by the yeast subtelomeric y' element, is induced in survivors defective for telomerase', *Journal of Biological Chemistry*. doi: 10.1074/jbc.273.50.33360.
Yim, G., Wang, H. H. and Davies, J. (2007) 'Antibiotics as signalling molecules', *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi: 10.1098/rstb.2007.2044.
Yingqing, M., Varoglu, M. and Sherman, D. H. (1999) 'Molecular characterization and analysis of the biosynthetic gene cluster for the antitumor antibiotic mitomycin C from Streptomyces lavendulae NRRL 2564', *Chemistry and Biology*. doi: 10.1016/S1074-5521(99)80040-4.
You, K. M., Rosenfield, C. and Knipple, D. C. (2003) 'Ethanol tolerance in the yeast Saccharomyces cerevisiae is dependent on cellular oleic acid content', *Applied and environmental microbiology*. doi: 10.1128/AEM.69.3.1499.
Yu, L. P., Wu, F. Q. and Chen, G. Q. (2019) 'Next-Generation Industrial Biotechnology-Transforming the Current Industrial Biotechnology into Competitive Processes', *Biotechnology Journal*. doi: 10.1002/biot.201800437.
Zeng, J. *et al.* (2020) 'Causalcall: Nanopore Basecalling Using a Temporal Convolutional Network', *Frontiers in Genetics*. doi: 10.3389/fgene.2019.01332.
Zhu, B. *et al.* (2020) 'Large-Scale Cultivation of Spirulina for Biological CO2 Mitigation in Open Raceway Ponds Using Purified CO2 From a Coal Chemical Flue Gas', *Frontiers in Bioengineering and Biotechnology*. doi: 10.3389/fbioe.2019.00441.
Zhulin, I. B. (2016) 'Classic spotlight: 16S rRNA redefines microbiology', *Journal of Bacteriology*. doi: 10.1128/JB.00616-16.
Ziemann, M. (2016) 'Accuracy, speed and error tolerance of short DNA sequence aligners', *bioRxiv*. doi: 10.1101/053686.
Zvyagintsev, D. G. (1994) *Vertical distribution of microbial communities in soils*, *Beyond the Biomass*.
Van Zyl, W. H., Bloom, M. and Viktor, M. J. (2012) 'Engineering yeasts for raw starch conversion', *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-012-4248-0.

Appendix 1.

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
1	172768	del	AT	A	A:50 AT:0	CDS	-	691/753	231/250	frameshift_variant c.691delA p.Ile231fs	TETR0P00860	hypothetical protein
1	223401	complex	ATTATCAG	TCCATTTGT	TCCATTTGT:20 ATTATCAG:0	CDS	-	513/516	169/171	frameshift_variant&missense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223426	ins	C	CT	CT:30 C:0	CDS	-	487/516	163/171	frameshift_variant c.487dupA p.Arg163fs	TETR0P01050	hypothetical protein
1	223451	ins	T	TG	TG:30 T:0	CDS	-	462/516	154/171	frameshift_variant c.462dupC p.Lys155fs	TETR0P01050	hypothetical protein
1	223459	del	CG	C	C:32 CG:0	CDS	-	454/516	152/171	frameshift_variant c.454delC p.Arg152fs	TETR0P01050	hypothetical protein
1	223549	ins	G	GGTA	GGTA:30 G:2	CDS	-	364/516	122/171	conservative_inframe_insertion c.364_365insTAC p.Ala121_Pro122insLeu	TETR0P01050	hypothetical protein
1	223594	complex	GG	TA	TA:36 GG:0	CDS	-	320/516	107/171	missense_variant c.319_320delCCinsTA p.Pro107Tyr	TETR0P01050	hypothetical protein
1	223608	ins	T	TTC	TTC:36 T:0	CDS	-	305/516	102/171	frameshift_variant c.304_305dupGA p.Asp103fs	TETR0P01050	hypothetical protein
1	223674	ins	T	TG	TG:38 T:0	CDS	-	239/516	80/171	frameshift_variant c.239dupC p.Ala81fs	TETR0P01050	hypothetical protein
1	223776	del	TG	T	T:28 TG:0	CDS	-	137/516	46/171	frameshift_variant c.137delC p.Pro46fs	TETR0P01050	hypothetical protein
1	223790	complex	ATCGT	TCA	TCA:24 ATCGT:0	CDS	-	124/516	40/171	frameshift_variant&missense_variant c.120_124delACGATinsTGA p.Arg41fs	TETR0P01050	hypothetical protein
1	223892	snp	T	C	C:34 T:0	CDS	-	22/516	8/171	missense_variant c.22A>G p.Arg8Gly	TETR0P01050	hypothetical protein
1	223904	complex	GC	A	A:32 GC:0	CDS	-	10/516	3/171	frameshift_variant&synonymous_variant c.9_10delGCinsT p.Gln3fs	TETR0P01050	hypothetical protein
3	2348	snp	T	C	C:156 T:4	CDS	-	133/363	45/120	missense_variant c.133A>G p.Thr45Ala	TETR0N00100	hypothetical protein
3	40270	snp	T	C	C:12 T:0	CDS	-	1539/4551	513/1516	synonymous_variant c.1539A>G p.Gln513Gln	TETR0N00220	hypothetical protein
3	302753	del	GA	G	G:68 GA:0	CDS	+	250/255	84/84	frameshift_variant c.250delA p.Met84fs	TETR0N01600	hypothetical protein
3	304736	del	CA	C	C:116 CA:0	CDS	+	1505/1716	502/571	frameshift_variant c.1505delA p.Asn502fs	TETR0N01620	hypothetical protein
3	304759	del	GA	G	G:128 GA:0	CDS	+	1527/1716	509/571	frameshift_variant c.1527delA p.Gly510fs	TETR0N01620	hypothetical protein
3	305286	del	TG	T	T:70 TG:0	CDS	+	414/423	138/140	frameshift_variant c.414delG p.His139fs	TETR0N01630	hypothetical protein
3	307409	del	TA	T	T:90 TA:0	CDS	-	377/384	126/127	frameshift_variant c.377delT p.Leu126fs	TETR0N01650	hypothetical protein
3	322191	del	GA	G	G:88 GA:0	CDS	+	100/1827	34/608	frameshift_variant c.100delA p.Thr34fs	TETR0N01730	hypothetical protein
3	323906	ins	G	GAA	GAA:82 G:0	CDS	+	1814/1827	605/608	frameshift_variant c.1812_1813dupAA p.Arg605fs	TETR0N01730	hypothetical protein
7	1598	del	TA	T	T:116 TA:0	CDS	+	14/1137	5/378	frameshift_variant c.14delA p.Asn5fs	TETR0C00100	hypothetical protein
7	6988	del	ACC	A	A:80 ACC:0	CDS	-	731/786	244/261	frameshift_variant c.730_731delGG p.Gly244fs	TETR0C00120	hypothetical protein
7	969667	del	GC	G	G:32 GC:0	CDS	-	378/387	126/128	frameshift_variant c.378delG p.Arg127fs	TETR0C04650	hypothetical protein
7	1056735	del	CT	C	C:90 CT:0	CDS	+	459/477	153/158	frameshift_variant c.459delT p.Phe153fs	TETR0C05090	hypothetical protein
9	421863	del	TA	T	T:82 TA:0	CDS	-	3768/3774	1256/1257	frameshift_variant c.3768delT p.Phe1256fs	TETR0M02150	hypothetical protein
9	427933	complex	TCCCT	CCC	CCC:38 TCCCT:2	CDS	+	30/474	10/157	frameshift_variant&synonymous_variant c.30_34delTCCCTinsCCC p.Phe12fs	TETR0M02170	hypothetical protein
9	428104	snp	T	A	A:10 T:0	CDS	+	201/474	67/157	synonymous_variant c.201T>A p.Ser67Ser	TETR0M02170	hypothetical protein
9	428119	snp	C	T	T:10 C:0	CDS	+	216/474	72/157	synonymous_variant c.216C>T p.His72His	TETR0M02170	hypothetical protein
9	428125	complex	CC	TG	TG:10 CC:0	CDS	+	222/474	74/157	missense_variant c.222_223delCCinsTG p.Gln75Glu	TETR0M02170	hypothetical protein
9	428134	snp	C	T	T:10 C:0	CDS	+	231/474	77/157	synonymous_variant c.231C>T p.Tyr77Tyr	TETR0M02170	hypothetical protein
9	428177	snp	G	A	A:16 G:0	CDS	+	274/474	92/157	missense_variant c.274G>A p.Val92Ile	TETR0M02170	hypothetical protein
9	428188	snp	C	T	T:18 C:0	CDS	+	285/474	95/157	synonymous_variant c.285C>T p.Ala95Ala	TETR0M02170	hypothetical protein
9	433156	del	CA	C	C:116 CA:0	CDS	+	830/1041	277/346	frameshift_variant c.830delA p.Asn277fs	TETR0M02200	hypothetical protein
9	433179	del	GA	G	G:128 GA:0	CDS	+	852/1041	284/346	frameshift_variant c.852delA p.Gly285fs	TETR0M02200	hypothetical protein
9	434306	del	TG	T	T:80 TG:0	CDS	+	861/882	287/293	frameshift_variant c.861delG p.Pro288fs	TETR0M02210	hypothetical protein
9	434442	del	GA	G	G:86 GA:0	CDS	+	29/1896	10/631	frameshift_variant c.29delA p.Lys10fs	TETR0M02220	hypothetical protein
10	5556	del	CA	C	C:18 CA:0	CDS	-	105/123	35/40	frameshift_variant c.105delT p.Phe35fs	TETR0I00120	hypothetical protein
10	14458	del	GA	G	G:94 GA:0	CDS	+	655/669	219/222	frameshift_variant c.655delA p.Ser219fs	TETR0I00160	hypothetical protein
10	19770	del	CA	C	C:50 CA:0	CDS	+	297/300	99/99	frameshift_variant c.297delA p.Lys99fs	TETR0I00190	hypothetical protein
10	24271	del	AT	A	A:60 AT:0	CDS	-	1196/1200	399/399	frameshift_variant c.1196delA p.Asn399fs	TETR0I00220	hypothetical protein
10	411768	del	TA	T	T:126 TA:0	CDS	+	48/51	16/16	frameshift_variant c.48delA p.Lys16fs	TETR0I02010	hypothetical protein
10	413050	snp	T	C	C:130 T:10	CDS	+	921/1608	307/535	synonymous_variant c.921T>C p.Tyr307Tyr	TETR0I02020	hypothetical protein
10	413354	snp	A	G	G:128 A:8	CDS	+	1225/1608	409/535	missense_variant c.1225A>G p.Lys409Glu	TETR0I02020	hypothetical protein
10	413669	del	GA	G	G:144 GA:0	CDS	+	1546/1608	516/535	frameshift_variant c.1546delA p.Ile516fs	TETR0I02020	hypothetical protein

Supp. Table 1.1 SNP details from snippy of strain 93 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
11	579211	ins	G	GT	GT:96 G:0	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
11	582417	del	GT	G	G:88 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
11	668463	del	AT	A	A:72 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
12	527213	del	AT	A	A:58 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
12	532232	ins	A	ATG	ATG:24 A:0	CDS	-	1/468	1/155	frameshift_variant&start_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein
12	371575	del	GA	G	G:92 GA:0	CDS	+	3351/3366	1117/1121	frameshift_variant c.3351delA p.Ala1118fs	TETR0B01760	hypothetical protein
12	374765	del	TG	T	T:124 TG:0	CDS	-	58/2388	20/795	frameshift_variant c.58delC p.His20fs	TETR0B01780	hypothetical protein
12	644064	ins	A	AT	AT:98 A:0	CDS	-	687/699	229/232	frameshift_variant c.687dupA p.Tyr230fs	TETR0B02920	hypothetical protein
12	822868	snp	A	G	G:46 A:0	CDS	+	1173/1434	391/477	synonymous_variant c.1173A>G p.Leu391Leu	TETR0B03730	hypothetical protein
12	823019	snp	A	G	G:38 A:0	CDS	+	1324/1434	442/477	missense_variant c.1324A>G p.Thr442Ala	TETR0B03730	hypothetical protein
12	824645	del	CT	C	C:40 CT:0	CDS	+	338/399	113/132	frameshift_variant c.338delT p.Leu113fs	TETR0B03740	hypothetical protein
12	835591	snp	A	G	G:44 A:0	CDS	+	762/942	254/313	synonymous_variant c.762A>G p.Leu254Leu	TETR0B03780	hypothetical protein
12	835702	complex	AGA	G	G:46 AGA:0	CDS	+	873/942	291/313	frameshift_variant&synonymous_variant c.873_875delAGAinsG p.Glu292fs	TETR0B03780	hypothetical protein
12	840158	del	AC	A	A:46 AC:0	CDS	-	561/645	187/214	frameshift_variant c.561delG p.Cys188fs	TETR0B03800	hypothetical protein
12	841416	snp	A	T	T:34 A:0	CDS	-	764/897	255/298	missense_variant c.764T>A p.Ile255Lys	TETR0B03810	hypothetical protein
13	47854	del	AT	A	A:34 AT:0	CDS	+	930/993	310/330	frameshift_variant c.930delT p.Phe310fs	TETR0E00320	hypothetical protein
13	320774	del	AC	A	A:28 AC:0	CDS	-	2235/2238	745/745	frameshift_variant c.2235delG p.Ter746fs	TETR0E01670	hypothetical protein
14	51012	del	AC	A	A:56 AC:0	CDS	+	351/372	117/123	frameshift_variant c.351delC p.Ile118fs	TETR0H00300	hypothetical protein
14	93111	snp	T	A	A:44 T:0	CDS	+	421/714	141/237	missense_variant c.421T>A p.Phe141Ile	TETR0H00450	hypothetical protein
14	108617	snp	A	G	G:38 A:0	CDS	+	838/1548	280/515	missense_variant c.838A>G p.Lys280Glu	TETR0H00510	hypothetical protein
14	721439	del	GT	G	G:88 GT:0	CDS	-	100/2190	34/729	frameshift_variant c.100delA p.Thr34fs	TETR0H03630	hypothetical protein
14	811774	del	GA	G	G:20 GA:0	CDS	-	38/747	13/248	frameshift_variant c.38delT p.Phe13fs	TETR0H04110	hypothetical protein
6	1084	del	CT	C	C:100 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
6	12603	del	GA	G	G:98 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
6	17063	del	AT	A	A:102 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
6	19889	del	AT	A	A:106 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
6	1716	ins	T	TC	TC:96 T:0	CDS	-	1057/1152	353/383	frameshift_variant c.1057dupG p.Glu353fs	TETR0D00110	hypothetical protein
6	2738	del	GT	G	G:84 GT:0	CDS	-	35/1152	12/383	frameshift_variant c.35delA p.Asn12fs	TETR0D00110	hypothetical protein
6	5820	complex	AT	C	C:68 AT:0	CDS	-	176/198	59/65	frameshift_variant&missense_variant c.175_176delATinsG p.Ile59fs	TETR0D00120	hypothetical protein
6	5940	snp	C	T	T:70 C:0	CDS	-	56/198	19/65	missense_variant c.56G>A p.Arg19Lys	TETR0D00120	hypothetical protein

Supp. Table 1.1 cont. SNP details from snippy of strain 93 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
1	172768	del	AT	A	A:50 AT:0	CDS	-	691/753	231/250	frameshift_variant c.691delA p.Ile231fs	TETR0P00860	hypothetical protein
1	223401	complex	ATTATCAG	TCCATTTGT	TCCATTTGT:20 ATTATCAG:0	CDS	-	513/516	169/171	frameshift_variant&missense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223426	ins	C	CT	CT:30 C:0	CDS	-	487/516	163/171	frameshift_variant c.487dupA p.Arg163fs	TETR0P01050	hypothetical protein
1	223451	ins	T	TG	TG:30 T:0	CDS	-	462/516	154/171	frameshift_variant c.462dupC p.Lys155fs	TETR0P01050	hypothetical protein
1	223459	del	CG	C	C:32 CG:0	CDS	-	454/516	152/171	frameshift_variant c.454delC p.Arg152fs	TETR0P01050	hypothetical protein
1	223549	ins	G	GGTA	GGTA:30 G:2	CDS	-	364/516	122/171	conservative_inframe_insertion c.364_365insTAC p.Ala121_Pro122insLeu	TETR0P01050	hypothetical protein
1	223594	complex	GG	TA	TA:36 GG:0	CDS	-	320/516	107/171	missense_variant c.319_320delCCinsTA p.Pro107Tyr	TETR0P01050	hypothetical protein
1	223608	ins	T	TTC	TTC:36 T:0	CDS	-	305/516	102/171	frameshift_variant c.304_305dupGA p.Asp103fs	TETR0P01050	hypothetical protein
1	223674	ins	T	TG	TG:38 T:0	CDS	-	239/516	80/171	frameshift_variant c.239dupC p.Ala81fs	TETR0P01050	hypothetical protein
1	223776	del	TG	T	T:28 TG:0	CDS	-	137/516	46/171	frameshift_variant c.137delC p.Pro46fs	TETR0P01050	hypothetical protein
1	223790	complex	ATCGT	TCA	TCA:24 ATCGT:0	CDS	-	124/516	40/171	frameshift_variant&missense_variant c.120_124delACGATinsTGA p.Arg41fs	TETR0P01050	hypothetical protein
1	223892	snp	T	C	C:34 T:0	CDS	-	22/516	8/171	missense_variant c.22A>G p.Arg8Gly	TETR0P01050	hypothetical protein
1	223904	complex	GC	A	A:32 GC:0	CDS	-	10/516	3/171	frameshift_variant&synonymous_variant c.9_10delGCinsT p.Gln3fs	TETR0P01050	hypothetical protein
3	2348	snp	T	C	C:156 T:4	CDS	-	133/363	45/120	missense_variant c.133A>G p.Thr45Ala	TETR0N00100	hypothetical protein
3	40270	snp	T	C	C:12 T:0	CDS	-	1539/4551	513/1516	synonymous_variant c.1539A>G p.Gln513Gln	TETR0N00220	hypothetical protein
3	302753	del	GA	G	G:68 GA:0	CDS	+	250/255	84/84	frameshift_variant c.250delA p.Met84fs	TETR0N01600	hypothetical protein
3	304736	del	CA	C	C:116 CA:0	CDS	+	1505/1716	502/571	frameshift_variant c.1505delA p.Asn502fs	TETR0N01620	hypothetical protein
3	304759	del	GA	G	G:128 GA:0	CDS	+	1527/1716	509/571	frameshift_variant c.1527delA p.Gly510fs	TETR0N01620	hypothetical protein
3	305286	del	TG	T	T:70 TG:0	CDS	+	414/423	138/140	frameshift_variant c.414delG p.His139fs	TETR0N01630	hypothetical protein
3	307409	del	TA	T	T:90 TA:0	CDS	-	377/384	126/127	frameshift_variant c.377delT p.Leu126fs	TETR0N01650	hypothetical protein
3	322191	del	GA	G	G:88 GA:0	CDS	+	100/1827	34/608	frameshift_variant c.100delA p.Thr34fs	TETR0N01730	hypothetical protein
3	323906	ins	G	GAA	GAA:82 G:0	CDS	+	1814/1827	605/608	frameshift_variant c.1812_1813dupAA p.Arg605fs	TETR0N01730	hypothetical protein
4	949016	ins	T	TG	TG:112 T:0	CDS	-	867/915	289/304	frameshift_variant c.867dupC p.Lys290fs	TETR0A04760	hypothetical protein
7	1598	del	TA	T	T:116 TA:0	CDS	+	14/1137	5/378	frameshift_variant c.14delA p.Asn5fs	TETR0C00100	hypothetical protein
7	6988	del	ACC	A	A:80 ACC:0	CDS	-	731/786	244/261	frameshift_variant c.730_731delGG p.Gly244fs	TETR0C00120	hypothetical protein
7	969667	del	GC	G	G:32 GC:0	CDS	-	378/387	126/128	frameshift_variant c.378delG p.Arg127fs	TETR0C04650	hypothetical protein
7	1056735	del	CT	C	C:90 CT:0	CDS	+	459/477	153/158	frameshift_variant c.459delT p.Phe153fs	TETR0C05090	hypothetical protein
9	421863	del	TA	T	T:82 TA:0	CDS	-	3768/3774	1256/1257	frameshift_variant c.3768delT p.Phe1256fs	TETR0M02150	hypothetical protein
9	427933	complex	TCCCT	CCC	CCC:38 TCCCT:2	CDS	+	30/474	10/157	frameshift_variant&synonymous_variant c.30_34delTCCCTinsCCC p.Phe12fs	TETR0M02170	hypothetical protein
9	428104	snp	T	A	A:10 T:0	CDS	+	201/474	67/157	synonymous_variant c.201T>A p.Ser67Ser	TETR0M02170	hypothetical protein
9	428119	snp	C	T	T:10 C:0	CDS	+	216/474	72/157	synonymous_variant c.216C>T p.His72His	TETR0M02170	hypothetical protein
9	428125	complex	CC	TG	TG:10 CC:0	CDS	+	222/474	74/157	missense_variant c.222_223delCCinsTG p.Gln75Glu	TETR0M02170	hypothetical protein
9	428134	snp	C	T	T:10 C:0	CDS	+	231/474	77/157	synonymous_variant c.231C>T p.Tyr77Tyr	TETR0M02170	hypothetical protein
9	428177	snp	G	A	A:16 G:0	CDS	+	274/474	92/157	missense_variant c.274G>A p.Val92Ile	TETR0M02170	hypothetical protein
9	428188	snp	C	T	T:18 C:0	CDS	+	285/474	95/157	synonymous_variant c.285C>T p.Ala95Ala	TETR0M02170	hypothetical protein
9	433156	del	CA	C	C:116 CA:0	CDS	+	830/1041	277/346	frameshift_variant c.830delA p.Asn277fs	TETR0M02200	hypothetical protein
9	433179	del	GA	G	G:128 GA:0	CDS	+	852/1041	284/346	frameshift_variant c.852delA p.Gly285fs	TETR0M02200	hypothetical protein
9	434306	del	TG	T	T:80 TG:0	CDS	+	861/882	287/293	frameshift_variant c.861delG p.Pro288fs	TETR0M02210	hypothetical protein
9	434442	del	GA	G	G:86 GA:0	CDS	+	29/1896	10/631	frameshift_variant c.29delA p.Lys10fs	TETR0M02220	hypothetical protein
10	5556	del	CA	C	C:18 CA:0	CDS	-	105/123	35/40	frameshift_variant c.105delT p.Phe35fs	TETR0I00120	hypothetical protein
10	14458	del	GA	G	G:94 GA:0	CDS	+	655/669	219/222	frameshift_variant c.655delA p.Ser219fs	TETR0I00160	hypothetical protein
10	19770	del	CA	C	C:50 CA:0	CDS	+	297/300	99/99	frameshift_variant c.297delA p.Lys99fs	TETR0I00190	hypothetical protein
10	24271	del	AT	A	A:60 AT:0	CDS	-	1196/1200	399/399	frameshift_variant c.1196delA p.Asn399fs	TETR0I00220	hypothetical protein
10	411768	del	TA	T	T:126 TA:0	CDS	+	48/51	16/16	frameshift_variant c.48delA p.Lys16fs	TETR0I02010	hypothetical protein
10	413050	snp	T	C	C:130 T:10	CDS	+	921/1608	307/535	synonymous_variant c.921T>C p.Tyr307Tyr	TETR0I02020	hypothetical protein
10	413354	snp	A	G	G:128 A:8	CDS	+	1225/1608	409/535	missense_variant c.1225A>G p.Lys409Glu	TETR0I02020	hypothetical protein
10	413669	del	GA	G	G:144 GA:0	CDS	+	1546/1608	516/535	frameshift_variant c.1546delA p.Ile516fs	TETR0I02020	hypothetical protein

Supp. Table 1.2. SNP details from snippy of strain 94 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
11	579211	ins	G	GT	GT:96 G:0	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
11	582417	del	GT	G	G:88 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
11	668463	del	AT	A	A:72 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
12	527213	del	AT	A	A:58 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
12	532232	ins	A	ATG	ATG:24 A:0	CDS	-	1/468	1/155	frameshift_variant&start_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein
12	371575	del	GA	G	G:92 GA:0	CDS	+	3351/3366	1117/1121	frameshift_variant c.3351delA p.Ala1118fs	TETR0B01760	hypothetical protein
12	374765	del	TG	T	T:124 TG:0	CDS	-	58/2388	20/795	frameshift_variant c.58delC p.His20fs	TETR0B01780	hypothetical protein
12	644064	ins	A	AT	AT:98 A:0	CDS	-	687/699	229/232	frameshift_variant c.687dupA p.Tyr230fs	TETR0B02920	hypothetical protein
12	822868	snp	A	G	G:46 A:0	CDS	+	1173/1434	391/477	synonymous_variant c.1173A>G p.Leu391Leu	TETR0B03730	hypothetical protein
12	823019	snp	A	G	G:38 A:0	CDS	+	1324/1434	442/477	missense_variant c.1324A>G p.Thr442Ala	TETR0B03730	hypothetical protein
12	824645	del	CT	C	C:40 CT:0	CDS	+	338/399	113/132	frameshift_variant c.338delT p.Leu113fs	TETR0B03740	hypothetical protein
12	835591	snp	A	G	G:44 A:0	CDS	+	762/942	254/313	synonymous_variant c.762A>G p.Leu254Leu	TETR0B03780	hypothetical protein
12	835702	complex	AGA	G	G:46 AGA:0	CDS	+	873/942	291/313	frameshift_variant&synonymous_variant c.873_875delAGAinsG p.Glu292fs	TETR0B03780	hypothetical protein
12	840158	del	AC	A	A:46 AC:0	CDS	-	561/645	187/214	frameshift_variant c.561delG p.Cys188fs	TETR0B03800	hypothetical protein
12	841416	snp	A	T	T:34 A:0	CDS	-	764/897	255/298	missense_variant c.764T>A p.Ile255Lys	TETR0B03810	hypothetical protein
13	47854	del	AT	A	A:34 AT:0	CDS	+	930/993	310/330	frameshift_variant c.930delT p.Phe310fs	TETR0E00320	hypothetical protein
13	320774	del	AC	A	A:28 AC:0	CDS	-	2235/2238	745/745	frameshift_variant c.2235delG p.Ter746fs	TETR0E01670	hypothetical protein
14	51012	del	AC	A	A:56 AC:0	CDS	+	351/372	117/123	frameshift_variant c.351delC p.Ile118fs	TETR0H00300	hypothetical protein
14	93111	snp	T	A	A:44 T:0	CDS	+	421/714	141/237	missense_variant c.421T>A p.Phe141Ile	TETR0H00450	hypothetical protein
14	108617	snp	A	G	G:38 A:0	CDS	+	838/1548	280/515	missense_variant c.838A>G p.Lys280Glu	TETR0H00510	hypothetical protein
14	721439	del	GT	G	G:88 GT:0	CDS	-	100/2190	34/729	frameshift_variant c.100delA p.Thr34fs	TETR0H03630	hypothetical protein
14	811774	del	GA	G	G:20 GA:0	CDS	-	38/747	13/248	frameshift_variant c.38delT p.Phe13fs	TETR0H04110	hypothetical protein
6	1084	del	CT	C	C:100 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
6	12603	del	GA	G	G:98 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
6	17063	del	AT	A	A:102 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
6	19889	del	AT	A	A:106 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
6	1716	ins	T	TC	TC:96 T:0	CDS	-	1057/1152	353/383	frameshift_variant c.1057dupG p.Glu353fs	TETR0D00110	hypothetical protein
6	2738	del	GT	G	G:84 GT:0	CDS	-	35/1152	12/383	frameshift_variant c.35delA p.Asn12fs	TETR0D00110	hypothetical protein
6	5820	complex	AT	C	C:68 AT:0	CDS	-	176/198	59/65	frameshift_variant&missense_variant c.175_176delATinsG p.Ile59fs	TETR0D00120	hypothetical protein
6	5940	snp	C	T	T:70 C:0	CDS	-	56/198	19/65	missense_variant c.56G>A p.Arg19Lys	TETR0D00120	hypothetical protein

Supp. Table 1.2 cont. SNP details from snippy of strain 94 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
1	172768	del	AT	A	A:397 AT:0	CDS	-	691/753	231/250	frameshift_variant c.691delA p.Ile231fs	TETR0P00860	hypothetical protein
1	223401	complex	ATTATCAG	TCCATTTGT	TCCATTTGT:65 ATTATCAG:0	CDS	-	513/516	169/171	frameshift_variant&missense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223426	ins	C	CT	CT:83 C:0	CDS	-	487/516	163/171	frameshift_variant&missense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223451	ins	T	TG	TG:80 T:0	CDS	-	462/516	154/171	frameshift_variant c.462dupC p.Lys155fs	TETR0P01050	hypothetical protein
1	223459	del	CG	C	C:83 CG:0	CDS	-	454/516	152/171	frameshift_variant c.454delC p.Arg152fs	TETR0P01050	hypothetical protein
1	223549	ins	G	GGTA	GGTA:74 G:6	CDS	-	364/516	122/171	conservative_inframe_insertion c.364_365insTAC p.Ala121_Pro122insLeu	TETR0P01050	hypothetical protein
1	223594	mnp	GG	TA	TA:71 GG:0	CDS	-	320/516	107/171	missense_variant c.319_320delCCinsTA p.Pro107Tyr	TETR0P01050	hypothetical protein
1	223608	ins	T	TTC	TTC:70 T:3	CDS	-	305/516	102/171	frameshift_variant c.304_305dupGA p.Asp103fs	TETR0P01050	hypothetical protein
1	223674	ins	T	TG	TG:65 T:0	CDS	-	239/516	80/171	frameshift_variant c.239dupC p.Ala81fs	TETR0P01050	hypothetical protein
1	223776	del	TG	T	T:66 TG:0	CDS	-	137/516	46/171	frameshift_variant c.137delC p.Pro46fs	TETR0P01050	hypothetical protein
1	223790	complex	ATCGT	TCA	TCA:58 ATCGT:0	CDS	-	124/516	40/171	frameshift_variant&missense_variant c.120_124delACGATinsTGA p.Arg41fs	TETR0P01050	hypothetical protein
1	223892	snp	T	C	C:54 T:0	CDS	-	22/516	8/171	missense_variant c.22A>G p.Arg8Gly	TETR0P01050	hypothetical protein
1	223904	complex	GC	A	A:55 GC:0	CDS	-	10/516	3/171	frameshift_variant&synonymous_variant c.9_10delGCinsT p.Gln3fs	TETR0P01050	hypothetical protein
1	517333	del	CT	C	C:45 CT:0	CDS	-	2218/2241	740/746	frameshift_variant c.2218delA p.Arg740fs	TETR0F02660	hypothetical protein
3	2348	snp	T	C	C:680 T:63	CDS	-	133/363	45/120	missense_variant c.133A>G p.Thr45Ala	TETR0N00100	hypothetical protein
3	302753	del	GA	G	G:157 GA:0	CDS	+	250/255	84/84	frameshift_variant c.250delA p.Met84fs	TETR0N01600	hypothetical protein
3	304736	del	CA	C	C:161 CA:0	CDS	+	1505/1716	502/571	frameshift_variant c.1505delA p.Asn502fs	TETR0N01620	hypothetical protein
3	304759	del	GA	G	G:178 GA:0	CDS	+	1527/1716	509/571	frameshift_variant c.1527delA p.Gly510fs	TETR0N01620	hypothetical protein
3	305286	del	TG	T	T:170 TG:0	CDS	+	414/423	138/140	frameshift_variant c.414delG p.His139fs	TETR0N01630	hypothetical protein
3	306726	snp	T	C	C:21 T:0	CDS	+	515/852	172/283	missense_variant c.515T>C p.Leu172Ser	TETR0N01640	hypothetical protein
3	306761	snp	C	T	T:24 C:1	CDS	+	550/852	184/283	missense_variant c.550C>T p.Pro184Ser	TETR0N01640	hypothetical protein
3	307409	del	TA	T	T:168 TA:0	CDS	-	377/384	126/127	frameshift_variant c.377delT p.Leu126fs	TETR0N01650	hypothetical protein
3	322191	del	GA	G	G:245 GA:0	CDS	+	100/1827	34/608	frameshift_variant c.100delA p.Thr34fs	TETR0N01730	hypothetical protein
3	323906	ins	G	GAA	GAA:232 G:0	CDS	+	1814/1827	605/608	frameshift_variant c.1812_1813dupAA p.Arg605fs	TETR0N01730	hypothetical protein
7	949016	ins	T	TG	TG:462 T:2	CDS	-	867/915	289/304	frameshift_variant c.867dupC p.Lys290fs	TETR0A04760	hypothetical protein
7	1084	del	CT	C	C:176 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
7	12603	del	GA	G	G:238 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
7	17063	del	AT	A	A:244 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
9	19889	del	AT	A	A:265 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
9	421863	del	TA	T	T:167 TA:0	CDS	-	3768/3774	1256/1257	frameshift_variant c.3768delT p.Phe1256fs	TETR0M02150	hypothetical protein
9	428231	del	GC	G	G:187 GC:0	CDS	+	333/474	111/157	frameshift_variant c.333delC p.Arg112fs	TETR0M02170	hypothetical protein
9	433156	del	CA	C	C:161 CA:0	CDS	+	830/1041	277/346	frameshift_variant c.830delA p.Asn277fs	TETR0M02200	hypothetical protein
9	433179	del	GA	G	G:178 GA:0	CDS	+	852/1041	284/346	frameshift_variant c.852delA p.Gly285fs	TETR0M02200	hypothetical protein
9	434306	del	TG	T	T:159 TG:0	CDS	+	861/882	287/293	frameshift_variant c.861delG p.Pro288fs	TETR0M02210	hypothetical protein
10	434442	del	GA	G	G:173 GA:0	CDS	+	29/1896	10/631	frameshift_variant c.29delA p.Lys10fs	TETR0M02220	hypothetical protein
10	527213	del	AT	A	A:96 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
10	532232	ins	A	ATG	ATG:20 A:0	CDS	-	1/468	1/155	frameshift_variant&start_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein
10	579211	ins	G	GT	GT:287 G:2	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
10	582417	del	GT	G	G:246 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
10	654602	del	CT	C	C:190 CT:0	CDS	-	1171/1182	391/393	frameshift_variant c.1171delA p.Ser391fs	TETR0J03170	hypothetical protein
10	668463	del	AT	A	A:138 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
10	674862	complex	ATG	GTA	GTA:14 ATG:0	CDS	-	2/474	1/157	start_lost c.-1_2delCATinsTAC p.Met1?	TETR0J03260	hypothetical protein
10	5556	del	CA	C	C:116 CA:0	CDS	-	105/123	35/40	frameshift_variant c.105delT p.Phe35fs	TETR0I00120	hypothetical protein
10	5753	del	TA	T	T:18 TA:0	CDS	-	416/438	139/145	frameshift_variant c.416delT p.Leu139fs	TETR0I00130	hypothetical protein
10	14458	del	GA	G	G:224 GA:0	CDS	+	655/669	219/222	frameshift_variant c.655delA p.Ser219fs	TETR0I00160	hypothetical protein
10	19770	del	CA	C	C:124 CA:0	CDS	+	297/300	99/99	frameshift_variant c.297delA p.Lys99fs	TETR0I00190	hypothetical protein
10	24271	del	AT	A	A:97 AT:0	CDS	-	1196/1200	399/399	frameshift_variant c.1196delA p.Asn399fs	TETR0I00220	hypothetical protein
10	411768	del	TA	T	T:456 TA:0	CDS	+	48/51	16/16	frameshift_variant c.48delA p.Lys16fs	TETR0I02010	hypothetical protein
10	413669	del	GA	G	G:394 GA:0	CDS	+	1546/1608	516/535	frameshift_variant c.1546delA p.Ile516fs	TETR0I02020	hypothetical protein
10	535096	snp	T	G	G:53 T:0	CDS	+	1801/2457	601/818	missense_variant c.1801T>G p.Tyr601Asp	TETR0I02620	hypothetical protein
10	536858	del	AG	A	A:54 AG:0	CDS	+	593/738	198/245	frameshift_variant c.593delG p.Gly198fs	TETR0I02630	hypothetical protein
10	536886	snp	T	C	C:56 T:0	CDS	+	619/738	207/245	missense_variant c.619T>C p.Trp207Arg	TETR0I02630	hypothetical protein
10	536937	del	GA	G	G:56 GA:0	CDS	+	674/738	225/245	frameshift_variant c.674delA p.Lys225fs	TETR0I02630	hypothetical protein

Supp. Table 1.3 SNP details from snippy of strain 132 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
11	579211	ins	G	GT	GT:96 G:0	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
11	582417	del	GT	G	G:88 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
11	668463	del	AT	A	A:72 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
12	527213	del	AT	A	A:58 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
12	532232	ins	A	ATG	ATG:24 A:0	CDS	-	1/468	1/155	frameshift_variant&start_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein
12	371575	del	GA	G	G:92 GA:0	CDS	+	3351/3366	1117/1121	frameshift_variant c.3351delA p.Ala1118fs	TETR0B01760	hypothetical protein
12	374765	del	TG	T	T:124 TG:0	CDS	-	58/2388	20/795	frameshift_variant c.58delC p.His20fs	TETR0B01780	hypothetical protein
12	644064	ins	A	AT	AT:98 A:0	CDS	-	687/699	229/232	frameshift_variant c.687dupA p.Tyr230fs	TETR0B02920	hypothetical protein
12	822868	snp	A	G	G:46 A:0	CDS	+	1173/1434	391/477	synonymous_variant c.1173A>G p.Leu391Leu	TETR0B03730	hypothetical protein
12	823019	snp	A	G	G:38 A:0	CDS	+	1324/1434	442/477	missense_variant c.1324A>G p.Thr442Ala	TETR0B03730	hypothetical protein
12	824645	del	CT	C	C:40 CT:0	CDS	+	338/399	113/132	frameshift_variant c.338delT p.Leu113fs	TETR0B03740	hypothetical protein
12	835591	snp	A	G	G:44 A:0	CDS	+	762/942	254/313	synonymous_variant c.762A>G p.Leu254Leu	TETR0B03780	hypothetical protein
12	835702	complex	AGA	G	G:46 AGA:0	CDS	+	873/942	291/313	frameshift_variant&synonymous_variant c.873_875delAGAinsG p.Glu292fs	TETR0B03780	hypothetical protein
12	840158	del	AC	A	A:46 AC:0	CDS	-	561/645	187/214	frameshift_variant c.561delG p.Cys188fs	TETR0B03800	hypothetical protein
12	841416	snp	A	T	T:34 A:0	CDS	-	764/897	255/298	missense_variant c.764T>A p.Ile255Lys	TETR0B03810	hypothetical protein
13	47854	del	AT	A	A:34 AT:0	CDS	+	930/993	310/330	frameshift_variant c.930delT p.Phe310fs	TETR0E00320	hypothetical protein
13	320774	del	AC	A	A:28 AC:0	CDS	-	2235/2238	745/745	frameshift_variant c.2235delG p.Ter746fs	TETR0E01670	hypothetical protein
14	51012	del	AC	A	A:56 AC:0	CDS	+	351/372	117/123	frameshift_variant c.351delC p.Ile118fs	TETR0H00300	hypothetical protein
14	93111	snp	T	A	A:44 T:0	CDS	+	421/714	141/237	missense_variant c.421T>A p.Phe141Ile	TETR0H00450	hypothetical protein
14	108617	snp	A	G	G:38 A:0	CDS	+	838/1548	280/515	missense_variant c.838A>G p.Lys280Glu	TETR0H00510	hypothetical protein
14	721439	del	GT	G	G:88 GT:0	CDS	-	100/2190	34/729	frameshift_variant c.100delA p.Thr34fs	TETR0H03630	hypothetical protein
14	811774	del	GA	G	G:20 GA:0	CDS	-	38/747	13/248	frameshift_variant c.38delT p.Phe13fs	TETR0H04110	hypothetical protein
6	1084	del	CT	C	C:100 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
6	12603	del	GA	G	G:98 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
6	17063	del	AT	A	A:102 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
6	19889	del	AT	A	A:106 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
6	1716	ins	T	TC	TC:96 T:0	CDS	-	1057/1152	353/383	frameshift_variant c.1057dupG p.Glu353fs	TETR0D00110	hypothetical protein
6	2738	del	GT	G	G:84 GT:0	CDS	-	35/1152	12/383	frameshift_variant c.35delA p.Asn12fs	TETR0D00110	hypothetical protein
6	5820	complex	AT	C	C:68 AT:0	CDS	-	176/198	59/65	frameshift_variant&missense_variant c.175_176delATinsG p.Ile59fs	TETR0D00120	hypothetical protein
6	5940	snp	C	T	T:70 C:0	CDS	-	56/198	19/65	missense_variant c.56G>A p.Arg19Lys	TETR0D00120	hypothetical protein

Supp. Table 1.3 cont. SNP details from snippy of strain 132 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
1	172768	del	AT	A	A:50 AT:0	CDS	-	691/753	231/250	frameshift_variant c.691delA p.Ile231fs	TETR0P00860	hypothetical protein
1	223401	complex	ATTATCAG	TCCATTTGT	TCCATTTGT:20 ATTATCAG:0	CDS	-	513/516	169/171	frameshift_variant&missense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223426	ins	C	CT	CT:30 C:0	CDS	-	487/516	163/171	frameshift_variant c.487dupA p.Arg163fs	TETR0P01050	hypothetical protein
1	223451	ins	T	TG	TG:30 T:0	CDS	-	462/516	154/171	frameshift_variant c.462dupC p.Lys155fs	TETR0P01050	hypothetical protein
1	223459	del	CG	C	C:32 CG:0	CDS	-	454/516	152/171	frameshift_variant c.454delC p.Arg152fs	TETR0P01050	hypothetical protein
1	223549	ins	G	GGTA	GGTA:30 G:2	CDS	-	364/516	122/171	conservative_inframe_insertion c.364_365insTAC p.Ala121_Pro122insLeu	TETR0P01050	hypothetical protein
1	223594	complex	GG	TA	TA:36 GG:0	CDS	-	320/516	107/171	missense_variant c.319_320delCCinsTA p.Pro107Tyr	TETR0P01050	hypothetical protein
1	223608	ins	T	TTC	TTC:36 T:0	CDS	-	305/516	102/171	frameshift_variant c.304_305dupGA p.Asp103fs	TETR0P01050	hypothetical protein
1	223674	ins	T	TG	TG:38 T:0	CDS	-	239/516	80/171	frameshift_variant c.239dupC p.Ala81fs	TETR0P01050	hypothetical protein
1	223776	del	TG	T	T:28 TG:0	CDS	-	137/516	46/171	frameshift_variant c.137delC p.Pro46fs	TETR0P01050	hypothetical protein
1	223790	complex	ATCGT	TCA	TCA:24 ATCGT:0	CDS	-	124/516	40/171	frameshift_variant&missense_variant c.120_124delACGATinsTGA p.Arg41fs	TETR0P01050	hypothetical protein
1	223892	snp	T	C	C:34 T:0	CDS	-	22/516	8/171	missense_variant c.22A>G p.Arg8Gly	TETR0P01050	hypothetical protein
1	223904	complex	GC	A	A:32 GC:0	CDS	-	10/516	3/171	frameshift_variant&synonymous_variant c.9_10delGCinsT p.Gln3fs	TETR0P01050	hypothetical protein
3	2348	snp	T	C	C:156 T:4	CDS	-	133/363	45/120	missense_variant c.133A>G p.Thr45Ala	TETR0N00100	hypothetical protein
3	40270	snp	T	C	C:12 T:0	CDS	-	1539/4551	513/1516	synonymous_variant c.1539A>G p.Gln513Gln	TETR0N00220	hypothetical protein
3	302753	del	GA	G	G:68 GA:0	CDS	+	250/255	84/84	frameshift_variant c.250delA p.Met84fs	TETR0N01600	hypothetical protein
3	304736	del	CA	C	C:116 CA:0	CDS	+	1505/1716	502/571	frameshift_variant c.1505delA p.Asn502fs	TETR0N01620	hypothetical protein
3	304759	del	GA	G	G:128 GA:0	CDS	+	1527/1716	509/571	frameshift_variant c.1527delA p.Gly510fs	TETR0N01620	hypothetical protein
3	305286	del	TG	T	T:70 TG:0	CDS	+	414/423	138/140	frameshift_variant c.414delG p.His139fs	TETR0N01630	hypothetical protein
3	307409	del	TA	T	T:90 TA:0	CDS	-	377/384	126/127	frameshift_variant c.377delT p.Leu126fs	TETR0N01650	hypothetical protein
3	322191	del	GA	G	G:88 GA:0	CDS	+	100/1827	34/608	frameshift_variant c.100delA p.Thr34fs	TETR0N01730	hypothetical protein
3	323906	ins	G	GAA	GAA:82 G:0	CDS	+	1814/1827	605/608	frameshift_variant c.1812_1813dupAA p.Arg605fs	TETR0N01730	hypothetical protein
4	949016	ins	T	TG	TG:112 T:0	CDS	-	867/915	289/304	frameshift_variant c.867dupC p.Lys290fs	TETR0A04760	hypothetical protein
6	1084	del	CT	C	C:100 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
6	12603	del	GA	G	G:98 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
6	17063	del	AT	A	A:102 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
6	19889	del	AT	A	A:106 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
7	1598	del	TA	T	T:116 TA:0	CDS	+	14/1137	5/378	frameshift_variant c.14delA p.Asn5fs	TETR0C00100	hypothetical protein
7	6988	del	ACC	A	A:80 ACC:0	CDS	-	731/786	244/261	frameshift_variant c.730_731delGG p.Gly244fs	TETR0C00120	hypothetical protein
7	969667	del	GC	G	G:32 GC:0	CDS	-	378/387	126/128	frameshift_variant c.378delG p.Arg127fs	TETR0C04650	hypothetical protein
7	1056735	del	CT	C	C:90 CT:0	CDS	+	459/477	153/158	frameshift_variant c.459delT p.Phe153fs	TETR0C05090	hypothetical protein
9	421863	del	TA	T	T:82 TA:0	CDS	-	3768/3774	1256/1257	frameshift_variant c.3768delT p.Phe1256fs	TETR0M02150	hypothetical protein
9	427933	complex	TCCCT	CCC	CCC:38 TCCCT:2	CDS	+	30/474	10/157	frameshift_variant&synonymous_variant c.30_34delTCCCTinsCCC p.Phe12fs	TETR0M02170	hypothetical protein
9	428104	snp	T	A	A:10 T:0	CDS	+	201/474	67/157	synonymous_variant c.201T>A p.Ser67Ser	TETR0M02170	hypothetical protein
9	428119	snp	C	T	T:10 C:0	CDS	+	216/474	72/157	synonymous_variant c.216C>T p.His72His	TETR0M02170	hypothetical protein
9	428125	complex	CC	TG	TG:10 CC:0	CDS	+	222/474	74/157	missense_variant c.222_223delCCinsTG p.Gln75Glu	TETR0M02170	hypothetical protein
9	428134	snp	C	T	T:10 C:0	CDS	+	231/474	77/157	synonymous_variant c.231C>T p.Tyr77Tyr	TETR0M02170	hypothetical protein
9	428177	snp	G	A	A:16 G:0	CDS	+	274/474	92/157	missense_variant c.274G>A p.Val92Ile	TETR0M02170	hypothetical protein

Supp. Table 1.4. SNP details from snippy of strain 133 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
9	428188	snp	C	T	T:18 C:0	CDS	+	285/474	95/157	synonymous_variant c.285C>T p.Ala95Ala	TETR0M02170	hypothetical protein
9	433156	del	CA	C	C:116 CA:0	CDS	+	830/1041	277/346	frameshift_variant c.830delA p.Asn277fs	TETR0M02200	hypothetical protein
9	433179	del	GA	G	G:128 GA:0	CDS	+	852/1041	284/346	frameshift_variant c.852delA p.Gly285fs	TETR0M02200	hypothetical protein
9	434306	del	TG	T	T:80 TG:0	CDS	+	861/882	287/293	frameshift_variant c.861delG p.Pro288fs	TETR0M02210	hypothetical protein
9	434442	del	GA	G	G:86 GA:0	CDS	+	29/1896	10/631	frameshift_variant c.29delA p.Lys10fs	TETR0M02220	hypothetical protein
10	5556	del	CA	C	C:18 CA:0	CDS	-	105/123	35/40	frameshift_variant c.105delT p.Phe35fs	TETR0I00120	hypothetical protein
10	14458	del	GA	G	G:94 GA:0	CDS	+	655/669	219/222	frameshift_variant c.655delA p.Ser219fs	TETR0I00160	hypothetical protein
10	19770	del	CA	C	C:50 CA:0	CDS	+	297/300	99/99	frameshift_variant c.297delA p.Lys99fs	TETR0I00190	hypothetical protein
10	24271	del	AT	A	A:60 AT:0	CDS	-	1196/1200	399/399	frameshift_variant c.1196delA p.Asn399fs	TETR0I00220	hypothetical protein
10	411768	del	TA	T	T:126 TA:0	CDS	+	48/51	16/16	frameshift_variant c.48delA p.Lys16fs	TETR0I02010	hypothetical protein
10	413050	snp	T	C	C:130 T:10	CDS	+	921/1608	307/535	synonymous_variant c.921T>C p.Tyr307Tyr	TETR0I02020	hypothetical protein
10	413354	snp	A	G	G:128 A:8	CDS	+	1225/1608	409/535	missense_variant c.1225A>G p.Lys409Glu	TETR0I02020	hypothetical protein
10	413669	del	GA	G	G:144 GA:0	CDS	+	1546/1608	516/535	frameshift_variant c.1546delA p.Ile516fs	TETR0I02020	hypothetical protein
11	579211	ins	G	GT	GT:96 G:0	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
11	582417	del	GT	G	G:88 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
11	668463	del	AT	A	A:72 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
12	527213	del	AT	A	A:58 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
12	532232	ins	A	ATG	ATG:24 A:0	CDS	-	1/468	1/155	frameshift_variant&start_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein
12	371575	del	GA	G	G:92 GA:0	CDS	+	3351/3366	1117/1121	frameshift_variant c.3351delA p.Ala1118fs	TETR0B01760	hypothetical protein
12	374765	del	TG	T	T:124 TG:0	CDS	-	58/2388	20/795	frameshift_variant c.58delC p.His20fs	TETR0B01780	hypothetical protein
12	644064	ins	A	AT	AT:98 A:0	CDS	-	687/699	229/232	frameshift_variant c.687dupA p.Tyr230fs	TETR0B02920	hypothetical protein
12	822868	snp	A	G	G:46 A:0	CDS	+	1173/1434	391/477	synonymous_variant c.1173A>G p.Leu391Leu	TETR0B03730	hypothetical protein
12	823019	snp	A	G	G:38 A:0	CDS	+	1324/1434	442/477	missense_variant c.1324A>G p.Thr442Ala	TETR0B03730	hypothetical protein
12	824645	del	CT	C	C:40 CT:0	CDS	+	338/399	113/132	frameshift_variant c.338delT p.Leu113fs	TETR0B03740	hypothetical protein
12	835591	snp	A	G	G:44 A:0	CDS	+	762/942	254/313	synonymous_variant c.762A>G p.Leu254Leu	TETR0B03780	hypothetical protein
12	835702	complex	AGA	G	G:46 AGA:0	CDS	+	873/942	291/313	frameshift_variant&synonymous_variant c.873_875delAGAinsG p.Glu292fs	TETR0B03780	hypothetical protein
12	840158	del	AC	A	A:46 AC:0	CDS	-	561/645	187/214	frameshift_variant c.561delG p.Cys188fs	TETR0B03800	hypothetical protein
12	841416	snp	A	T	T:34 A:0	CDS	-	764/897	255/298	missense_variant c.764T>A p.Ile255Lys	TETR0B03810	hypothetical protein
13	47854	del	AT	A	A:34 AT:0	CDS	+	930/993	310/330	frameshift_variant c.930delT p.Phe310fs	TETR0E00320	hypothetical protein
13	320774	del	AC	A	A:28 AC:0	CDS	-	2235/2238	745/745	frameshift_variant c.2235delG p.Ter746fs	TETR0E01670	hypothetical protein
14	51012	del	AC	A	A:56 AC:0	CDS	+	351/372	117/123	frameshift_variant c.351delC p.Ile118fs	TETR0H00300	hypothetical protein
14	93111	snp	T	A	A:44 T:0	CDS	+	421/714	141/237	missense_variant c.421T>A p.Phe141Ile	TETR0H00450	hypothetical protein
14	108617	snp	A	G	G:38 A:0	CDS	+	838/1548	280/515	missense_variant c.838A>G p.Lys280Glu	TETR0H00510	hypothetical protein
14	721439	del	GT	G	G:88 GT:0	CDS	-	100/2190	34/729	frameshift_variant c.100delA p.Thr34fs	TETR0H03630	hypothetical protein
14	811774	del	GA	G	G:20 GA:0	CDS	-	38/747	13/248	frameshift_variant c.38delT p.Phe13fs	TETR0H04110	hypothetical protein

Supp. Table 1.4 cont. SNP details from snippy of strain 133 compared to Ethanol Red

Chromosome	n	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
1	172768	del	AT	A	A:50 AT:0	CDS	-	691/753	231/250	frameshift_variant c.691delA p.Ile231fs	TETR0P00860	hypothetical protein
1	223401	complex	ATTATCAG	TCCATTTGT	TCCATTTGT:20 ATTATCAG:0	CDS	-	513/516	169/171	ense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223426	ins	C	CT	CT:30 C:0	CDS	-	487/516	163/171	frameshift_variant c.487dupA p.Arg163fs	TETR0P01050	hypothetical protein
1	223451	ins	T	TG	TG:30 T:0	CDS	-	462/516	154/171	frameshift_variant c.462dupC p.Lys155fs	TETR0P01050	hypothetical protein
1	223459	del	CG	C	C:32 CG:0	CDS	-	454/516	152/171	frameshift_variant c.454delC p.Arg152fs	TETR0P01050	hypothetical protein
1	223549	ins	G	GGTA	GGTA:30 G:2	CDS	-	364/516	122/171	conservative_inframe_insertion c.364_365insTAC p.Ala121_Pro122insLeu	TETR0P01050	hypothetical protein
1	223594	complex	GG	TA	TA:36 GG:0	CDS	-	320/516	107/171	c.319_320delCCinsTA p.Pro107Tyr	TETR0P01050	hypothetical protein
1	223608	ins	T	TTC	TTC:36 T:0	CDS	-	305/516	102/171	c.304_305dupGA p.Asp103fs	TETR0P01050	hypothetical protein
1	223674	ins	T	TG	TG:38 T:0	CDS	-	239/516	80/171	frameshift_variant c.239dupC p.Ala81fs	TETR0P01050	hypothetical protein
1	223776	del	TG	T	T:28 TG:0	CDS	-	137/516	46/171	frameshift_variant c.137delC p.Pro46fs	TETR0P01050	hypothetical protein
1	223790	complex	ATCGT	TCA	TCA:24 ATCGT:0	CDS	-	124/516	40/171	ense_variant c.120_124delACGATinsTGA p.Arg41fs	TETR0P01050	hypothetical protein
1	223892	snp	T	C	C:34 T:0	CDS	-	22/516	8/171	missense_variant c.22A>G p.Arg8Gly	TETR0P01050	hypothetical protein
1	223904	complex	GC	A	A:32 GC:0	CDS	-	10/516	3/171	nymous_variant c.9_10delGCinsT p.Gln3fs	TETR0P01050	hypothetical protein
3	2348	snp	T	C	C:156 T:4	CDS	-	133/363	45/120	missense_variant c.133A>G p.Thr45Ala	TETR0N00100	hypothetical protein
3	40270	snp	T	C	C:12 T:0	CDS	-	1539/4551	513/1516	synonymous_variant c.1539A>G p.Gln513Gln	TETR0N00220	hypothetical protein
3	302753	del	GA	G	G:68 GA:0	CDS	+	250/255	84/84	frameshift_variant c.250delA p.Met84fs	TETR0N01600	hypothetical protein
3	304736	del	CA	C	C:116 CA:0	CDS	+	1505/1716	502/571	frameshift_variant c.1505delA p.Asn502fs	TETR0N01620	hypothetical protein
3	304759	del	GA	G	G:128 GA:0	CDS	+	1527/1716	509/571	frameshift_variant c.1527delA p.Gly510fs	TETR0N01620	hypothetical protein
3	305286	del	TG	T	T:70 TG:0	CDS	+	414/423	138/140	frameshift_variant c.414delG p.His139fs	TETR0N01630	hypothetical protein
3	307409	del	TA	T	T:90 TA:0	CDS	-	377/384	126/127	frameshift_variant c.377delT p.Leu126fs	TETR0N01650	hypothetical protein
3	322191	del	GA	G	G:88 GA:0	CDS	+	100/1827	34/608	frameshift_variant c.100delA p.Thr34fs	TETR0N01730	hypothetical protein
3	323906	ins	G	GAA	GAA:82 G:0	CDS	+	1814/1827	605/608	c.1812_1813dupAA p.Arg605fs	TETR0N01730	hypothetical protein
4	949016	ins	T	TG	TG:112 T:0	CDS	-	867/915	289/304	frameshift_variant c.867dupC p.Lys290fs	TETR0A04760	hypothetical protein
6	1084	del	CT	C	C:100 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
6	12603	del	GA	G	G:98 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
6	17063	del	AT	A	A:102 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
6	19889	del	AT	A	A:106 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
6	1716	ins	T	TC	TC:96 T:0	CDS	-	1057/1152	353/383	frameshift_variant c.1057dupG p.Glu353fs	TETR0D00110	hypothetical protein
6	2738	del	GT	G	G:84 GT:0	CDS	-	35/1152	12/383	frameshift_variant c.35delA p.Asn12fs	TETR0D00110	hypothetical protein
6	5820	complex	AT	C	C:68 AT:0	CDS	-	176/198	59/65	ense_variant c.175_176delATinsG	TETR0D00120	hypothetical protein
6	5940	snp	C	T	T:70 C:0	CDS	-	56/198	19/65	missense_variant c.56G>A p.Arg19Lys	TETR0D00120	hypothetical protein
7	1598	del	TA	T	T:116 TA:0	CDS	+	14/1137	5/378	frameshift_variant c.14delA p.Asn5fs	TETR0C00100	hypothetical protein
7	6988	del	ACC	A	A:80 ACC:0	CDS	-	731/786	244/261	c.730_731delGG p.Gly244fs	TETR0C00120	hypothetical protein
7	969667	del	GC	G	G:32 GC:0	CDS	-	378/387	126/128	frameshift_variant c.378delG p.Arg127fs	TETR0C04650	hypothetical protein
7	1056735	del	CT	C	C:90 CT:0	CDS	+	459/477	153/158	frameshift_variant c.459delT p.Phe153fs	TETR0C05090	hypothetical protein
8	527213	del	AT	A	A:58 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
8	532232	ins	A	ATG	ATG:24 A:0	CDS	-	1/468	1/155	_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein

Supp. Table 1.5 cont. SNP details from snippy of strain 134 compared to Ethanol Red

Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
9	421863	del	TA	T	T:82 TA:0	CDS	-	3768/3774	1256/1257	frameshift_variant c.3768delT p.Phe1256fs	TETR0M02150	hypothetical protein
9	427933	complex	TCCCT	CCC	CCC:38 TCCCT:2	CDS	+	30/474	10/157	frameshift_variant&synonymous_variant c.30_34delTCCCTinsCCC p.Phe12fs	TETR0M02170	hypothetical protein
9	428104	snp	T	A	A:10 T:0	CDS	+	201/474	67/157	synonymous_variant c.201T>A p.Ser67Ser	TETR0M02170	hypothetical protein
9	428119	snp	C	T	T:10 C:0	CDS	+	216/474	72/157	synonymous_variant c.216C>T p.His72His	TETR0M02170	hypothetical protein
9	428125	complex	CC	TG	TG:10 CC:0	CDS	+	222/474	74/157	missense_variant c.222_223delCCinsTG p.Gln75Glu	TETR0M02170	hypothetical protein
9	428134	snp	C	T	T:10 C:0	CDS	+	231/474	77/157	synonymous_variant c.231C>T p.Tyr77Tyr	TETR0M02170	hypothetical protein
9	428177	snp	G	A	A:16 G:0	CDS	+	274/474	92/157	missense_variant c.274G>A p.Val92Ile	TETR0M02170	hypothetical protein
9	428188	snp	C	T	T:18 C:0	CDS	+	285/474	95/157	synonymous_variant c.285C>T p.Ala95Ala	TETR0M02170	hypothetical protein
9	433156	del	CA	C	C:116 CA:0	CDS	+	830/1041	277/346	frameshift_variant c.830delA p.Asn277fs	TETR0M02200	hypothetical protein
9	433179	del	GA	G	G:128 GA:0	CDS	+	852/1041	284/346	frameshift_variant c.852delA p.Gly285fs	TETR0M02200	hypothetical protein
9	434306	del	TG	T	T:80 TG:0	CDS	+	861/882	287/293	frameshift_variant c.861delG p.Pro288fs	TETR0M02210	hypothetical protein
9	434442	del	GA	G	G:86 GA:0	CDS	+	29/1896	10/631	frameshift_variant c.29delA p.Lys10fs	TETR0M02220	hypothetical protein
10	5556	del	CA	C	C:18 CA:0	CDS	-	105/123	35/40	frameshift_variant c.105delT p.Phe35fs	TETR0I00120	hypothetical protein
10	14458	del	GA	G	G:94 GA:0	CDS	+	655/669	219/222	frameshift_variant c.655delA p.Ser219fs	TETR0I00160	hypothetical protein
10	19770	del	CA	C	C:50 CA:0	CDS	+	297/300	99/99	frameshift_variant c.297delA p.Lys99fs	TETR0I00190	hypothetical protein
10	24271	del	AT	A	A:60 AT:0	CDS	-	1196/1200	399/399	frameshift_variant c.1196delA p.Asn399fs	TETR0I00220	hypothetical protein
10	411768	del	TA	T	T:126 TA:0	CDS	+	48/51	16/16	frameshift_variant c.48delA p.Lys16fs	TETR0I02010	hypothetical protein
10	413050	snp	T	C	C:130 T:10	CDS	+	921/1608	307/535	synonymous_variant c.921T>C p.Tyr307Tyr	TETR0I02020	hypothetical protein
10	413354	snp	A	G	G:128 A:8	CDS	+	1225/1608	409/535	missense_variant c.1225A>G p.Lys409Glu	TETR0I02020	hypothetical protein
10	413669	del	GA	G	G:144 GA:0	CDS	+	1546/1608	516/535	frameshift_variant c.1546delA p.Ile516fs	TETR0I02020	hypothetical protein
11	579211	ins	G	GT	GT:96 G:0	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
11	582417	del	GT	G	G:88 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
11	668463	del	AT	A	A:72 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
13	47854	del	AT	A	A:34 AT:0	CDS	+	930/993	310/330	frameshift_variant c.930delT p.Phe310fs	TETR0E00320	hypothetical protein
13	320774	del	AC	A	A:28 AC:0	CDS	-	2235/2238	745/745	frameshift_variant c.2235delG p.Ter746fs	TETR0E01670	hypothetical protein
14	51012	del	AC	A	A:56 AC:0	CDS	+	351/372	117/123	frameshift_variant c.351delC p.Ile118fs	TETR0H00300	hypothetical protein
14	93111	snp	T	A	A:44 T:0	CDS	+	421/714	141/237	missense_variant c.421T>A p.Phe141Ile	TETR0H00450	hypothetical protein
14	108617	snp	A	G	G:38 A:0	CDS	+	838/1548	280/515	missense_variant c.838A>G p.Lys280Glu	TETR0H00510	hypothetical protein
14	721439	del	GT	G	G:88 GT:0	CDS	-	100/2190	34/729	frameshift_variant c.100delA p.Thr34fs	TETR0H03630	hypothetical protein
14	811774	del	GA	G	G:20 GA:0	CDS	-	38/747	13/248	frameshift_variant c.38delT p.Phe13fs	TETR0H04110	hypothetical protein

Supp. Table 1.5. SNP details from snippy of strain 134 compared to Ethanol Red

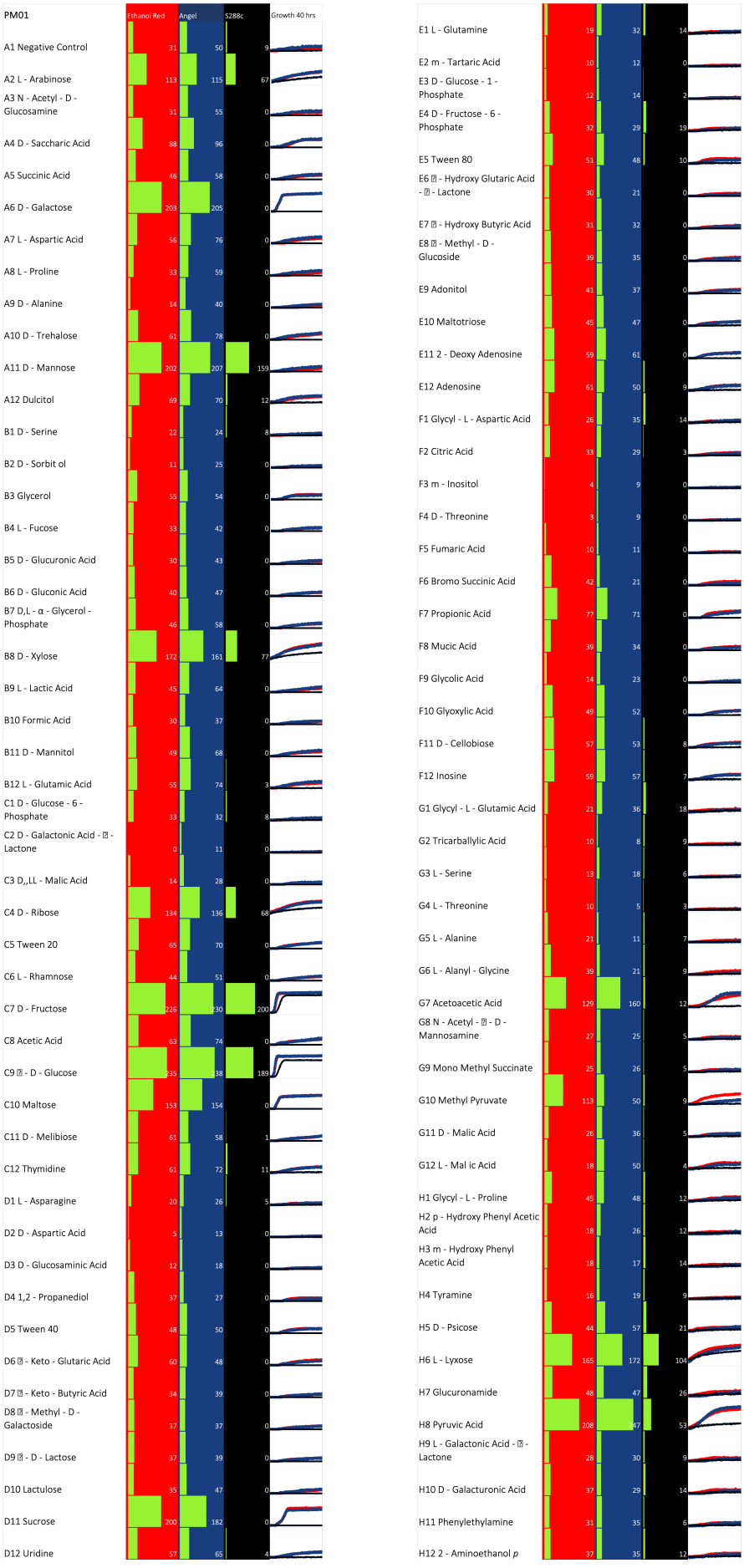
Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
1	172768	del	AT	A	A:50 AT:0	CDS	-	691/753	231/250	frameshift_variant c.691delA p.Ile231fs	TETR0P00860	hypothetical protein
1	223401	complex	ATTATCAG	TCCATTTGT	TCCATTTGT:20 ATTATCAG:0	CDS	-	513/516	169/171	frameshift_variant&missense_variant c.506_513delCTGATAATinsACAAATGGA p.Pro169fs	TETR0P01050	hypothetical protein
1	223426	ins	C	CT	CT:30 C:0	CDS	-	487/516	163/171	frameshift_variant c.487dupA p.Arg163fs	TETR0P01050	hypothetical protein
1	223451	ins	T	TG	TG:30 T:0	CDS	-	462/516	154/171	frameshift_variant c.462dupC p.Lys155fs	TETR0P01050	hypothetical protein
1	223459	del	CG	C	C:32 CG:0	CDS	-	454/516	152/171	frameshift_variant c.454delC p.Arg152fs	TETR0P01050	hypothetical protein
1	223549	ins	G	GGTA	GGTA:30 G:2	CDS	-	364/516	122/171	conservative_inframe_insertion c.364_365insTAC p.Ala121_Pro122insLeu	TETR0P01050	hypothetical protein
1	223594	complex	GG	TA	TA:36 GG:0	CDS	-	320/516	107/171	missense_variant c.319_320delCCinsTA p.Pro107Tyr	TETR0P01050	hypothetical protein
1	223608	ins	T	TTC	TTC:36 T:0	CDS	-	305/516	102/171	frameshift_variant c.304_305dupGA p.Asp103fs	TETR0P01050	hypothetical protein
1	223674	ins	T	TG	TG:38 T:0	CDS	-	239/516	80/171	frameshift_variant c.239dupC p.Ala81fs	TETR0P01050	hypothetical protein
1	223776	del	TG	T	T:28 TG:0	CDS	-	137/516	46/171	frameshift_variant c.137delC p.Pro46fs	TETR0P01050	hypothetical protein
1	223790	complex	ATCGT	TCA	TCA:24 ATCGT:0	CDS	-	124/516	40/171	frameshift_variant&missense_variant c.120_124delACGATinsTGA p.Arg41fs	TETR0P01050	hypothetical protein
1	223892	snp	T	C	C:34 T:0	CDS	-	22/516	8/171	missense_variant c.22A>G p.Arg8Gly	TETR0P01050	hypothetical protein
1	223904	complex	GC	A	A:32 GC:0	CDS	-	10/516	3/171	frameshift_variant&synonymous_variant c.9_10delGCinsT p.Gln3fs	TETR0P01050	hypothetical protein
3	2348	snp	T	C	C:156 T:4	CDS	-	133/363	45/120	missense_variant c.133A>G p.Thr45Ala	TETR0N00100	hypothetical protein
3	40270	snp	T	C	C:12 T:0	CDS	-	1539/4551	513/1516	synonymous_variant c.1539A>G p.Gln513Gln	TETR0N00220	hypothetical protein
3	302753	del	GA	G	G:68 GA:0	CDS	+	250/255	84/84	frameshift_variant c.250delA p.Met84fs	TETR0N01600	hypothetical protein
3	304736	del	CA	C	C:116 CA:0	CDS	+	1505/1716	502/571	frameshift_variant c.1505delA p.Asn502fs	TETR0N01620	hypothetical protein
3	304759	del	GA	G	G:128 GA:0	CDS	+	1527/1716	509/571	frameshift_variant c.1527delA p.Gly510fs	TETR0N01620	hypothetical protein
3	305286	del	TG	T	T:70 TG:0	CDS	+	414/423	138/140	frameshift_variant c.414delG p.His139fs	TETR0N01630	hypothetical protein
3	307409	del	TA	T	T:90 TA:0	CDS	-	377/384	126/127	frameshift_variant c.377delT p.Leu126fs	TETR0N01650	hypothetical protein
3	322191	del	GA	G	G:88 GA:0	CDS	+	100/1827	34/608	frameshift_variant c.100delA p.Thr34fs	TETR0N01730	hypothetical protein
3	323906	ins	G	GAA	GAA:82 G:0	CDS	+	1814/1827	605/608	frameshift_variant c.1812_1813dupAA p.Arg605fs	TETR0N01730	hypothetical protein
4	949016	ins	T	TG	TG:112 T:0	CDS	-	867/915	289/304	frameshift_variant c.867dupC p.Lys290fs	TETR0A04760	hypothetical protein
6	1084	del	CT	C	C:100 CT:0	CDS	-	1974/2109	658/702	frameshift_variant c.1974delA p.Val659fs	TETR0O00100	hypothetical protein
6	12603	del	GA	G	G:98 GA:0	CDS	+	34/3294	12/1097	frameshift_variant c.34delA p.Thr12fs	TETR0O00120	hypothetical protein
6	17063	del	AT	A	A:102 AT:0	CDS	-	1219/1239	407/412	frameshift_variant c.1219delA p.Ile407fs	TETR0O00130	hypothetical protein
6	19889	del	AT	A	A:106 AT:0	CDS	+	640/741	214/246	frameshift_variant c.640delT p.Ser214fs	TETR0O00140	hypothetical protein
6	1716	ins	T	TC	TC:96 T:0	CDS	-	1057/1152	353/383	frameshift_variant c.1057dupG p.Glu353fs	TETR0D00110	hypothetical protein
6	2738	del	GT	G	G:84 GT:0	CDS	-	35/1152	12/383	frameshift_variant c.35delA p.Asn12fs	TETR0D00110	hypothetical protein
6	5820	complex	AT	C	C:68 AT:0	CDS	-	176/198	59/65	frameshift_variant&missense_variant c.175_176delATinsG p.Ile59fs	TETR0D00120	hypothetical protein
6	5940	snp	C	T	T:70 C:0	CDS	-	56/198	19/65	missense_variant c.56G>A p.Arg19Lys	TETR0D00120	hypothetical protein
7	1598	del	TA	T	T:116 TA:0	CDS	+	14/1137	5/378	frameshift_variant c.14delA p.Asn5fs	TETR0C00100	hypothetical protein
7	6988	del	ACC	A	A:80 ACC:0	CDS	-	731/786	244/261	frameshift_variant c.730_731delGG p.Gly244fs	TETR0C00120	hypothetical protein
7	969667	del	GC	G	G:32 GC:0	CDS	-	378/387	126/128	frameshift_variant c.378delG p.Arg127fs	TETR0C04650	hypothetical protein
7	1056735	del	CT	C	C:90 CT:0	CDS	+	459/477	153/158	frameshift_variant c.459delT p.Phe153fs	TETR0C05090	hypothetical protein
9	421863	del	TA	T	T:82 TA:0	CDS	-	3768/3774	1256/1257	frameshift_variant c.3768delT p.Phe1256fs	TETR0M02150	hypothetical protein
9	427933	complex	TCCCT	CCC	CCC:38 TCCCT:2	CDS	+	30/474	10/157	frameshift_variant&synonymous_variant c.30_34delTCCCTinsCCC p.Phe12fs	TETR0M02170	hypothetical protein
9	428104	snp	T	A	A:10 T:0	CDS	+	201/474	67/157	synonymous_variant c.201T>A p.Ser67Ser	TETR0M02170	hypothetical protein

Supp. Table 1.6. SNP details from snippy of strain 135 compared to Ethanol Red

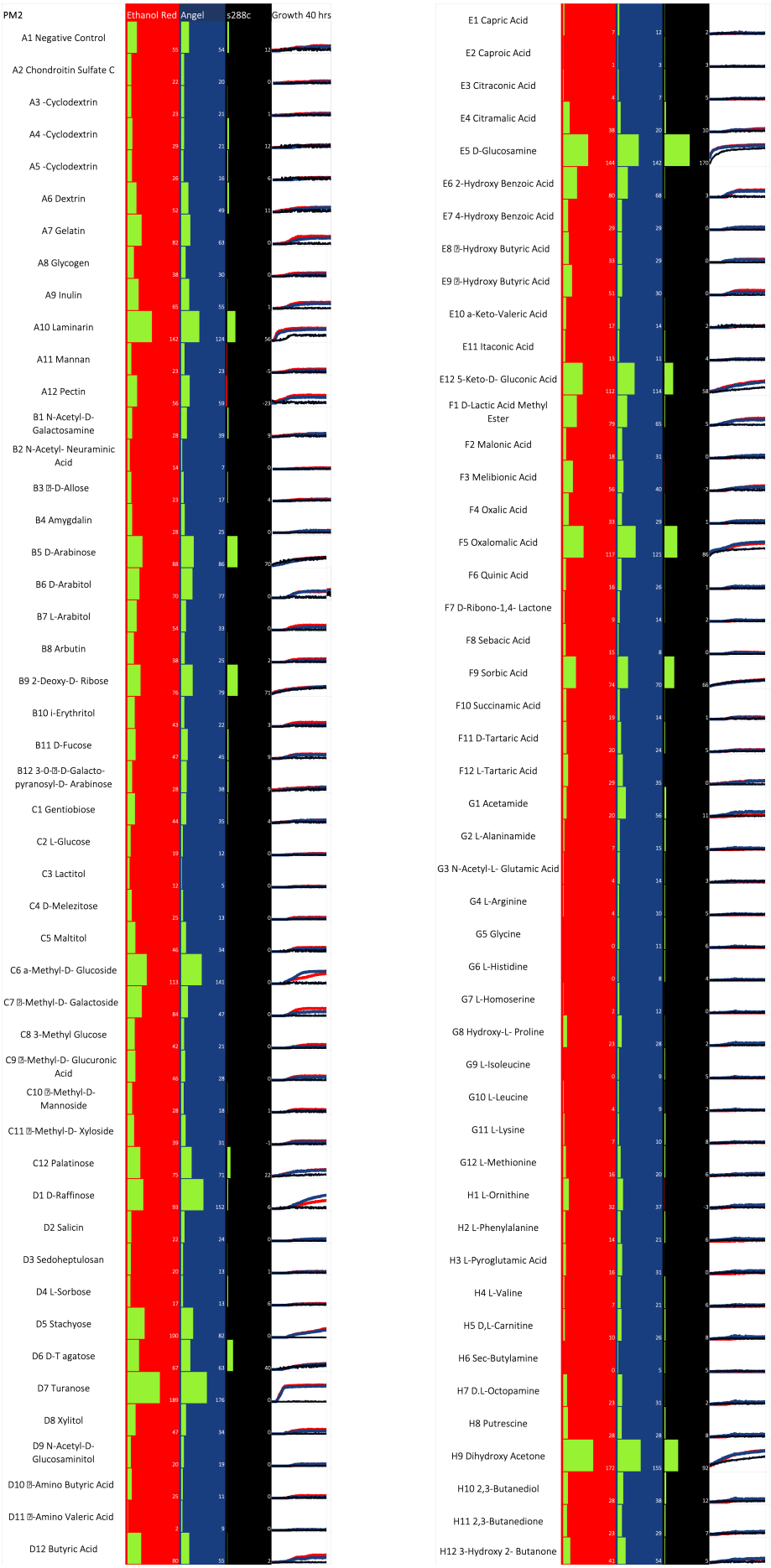
Chromosome	Scaffold_position	Type	Ref	ALT	Evidence	FType	STRAND	NT_Scaffold_position	AA_Scaffold_position	Effect	Gene	Product
9	428119	snp	C	T	T:10 C:0	CDS	+	216/474	72/157	synonymous_variant c.216C>T p.His72His	TETR0M02170	hypothetical protein
9	428125	complex	CC	TG	TG:10 CC:0	CDS	+	222/474	74/157	missense_variant c.222_223delCCinsTG p.Gln75Glu	TETR0M02170	hypothetical protein
9	428134	snp	C	T	T:10 C:0	CDS	+	231/474	77/157	synonymous_variant c.231C>T p.Tyr77Tyr	TETR0M02170	hypothetical protein
9	428177	snp	G	A	A:16 G:0	CDS	+	274/474	92/157	missense_variant c.274G>A p.Val92Ile	TETR0M02170	hypothetical protein
9	428188	snp	C	T	T:18 C:0	CDS	+	285/474	95/157	synonymous_variant c.285C>T p.Ala95Ala	TETR0M02170	hypothetical protein
9	433156	del	CA	C	C:116 CA:0	CDS	+	830/1041	277/346	frameshift_variant c.830delA p.Asn277fs	TETR0M02200	hypothetical protein
9	433179	del	GA	G	G:128 GA:0	CDS	+	852/1041	284/346	frameshift_variant c.852delA p.Gly285fs	TETR0M02200	hypothetical protein
9	434306	del	TG	T	T:80 TG:0	CDS	+	861/882	287/293	frameshift_variant c.861delG p.Pro288fs	TETR0M02210	hypothetical protein
9	434442	del	GA	G	G:86 GA:0	CDS	+	29/1896	10/631	frameshift_variant c.29delA p.Lys10fs	TETR0M02220	hypothetical protein
10	5556	del	CA	C	C:18 CA:0	CDS	-	105/123	35/40	frameshift_variant c.105delT p.Phe35fs	TETR0I00120	hypothetical protein
10	14458	del	GA	G	G:94 GA:0	CDS	+	655/669	219/222	frameshift_variant c.655delA p.Ser219fs	TETR0I00160	hypothetical protein
10	19770	del	CA	C	C:50 CA:0	CDS	+	297/300	99/99	frameshift_variant c.297delA p.Lys99fs	TETR0I00190	hypothetical protein
10	24271	del	AT	A	A:60 AT:0	CDS	-	1196/1200	399/399	frameshift_variant c.1196delA p.Asn399fs	TETR0I00220	hypothetical protein
10	411768	del	TA	T	T:126 TA:0	CDS	+	48/51	16/16	frameshift_variant c.48delA p.Lys16fs	TETR0I02010	hypothetical protein
10	413050	snp	T	C	C:130 T:10	CDS	+	921/1608	307/535	synonymous_variant c.921T>C p.Tyr307Tyr	TETR0I02020	hypothetical protein
10	413354	snp	A	G	G:128 A:8	CDS	+	1225/1608	409/535	missense_variant c.1225A>G p.Lys409Glu	TETR0I02020	hypothetical protein
10	413669	del	GA	G	G:144 GA:0	CDS	+	1546/1608	516/535	frameshift_variant c.1546delA p.Ile516fs	TETR0I02020	hypothetical protein
11	579211	ins	G	GT	GT:96 G:0	CDS	-	3306/3321	1102/1106	frameshift_variant c.3306dupA p.His1103fs	TETR0J02850	hypothetical protein
11	582417	del	GT	G	G:88 GT:0	CDS	-	100/3321	34/1106	frameshift_variant c.100delA p.Thr34fs	TETR0J02850	hypothetical protein
11	668463	del	AT	A	A:72 AT:0	CDS	+	465/471	155/156	frameshift_variant c.465delT p.Phe155fs	TETR0J03230	hypothetical protein
12	527213	del	AT	A	A:58 AT:0	CDS	+	77/1359	26/452	frameshift_variant c.77delT p.Leu26fs	TETR0L02670	hypothetical protein
12	532232	ins	A	ATG	ATG:24 A:0	CDS	-	1/468	1/155	frameshift_variant&start_lost c.1_2insCA p.Met1fs	TETR0L02700	hypothetical protein
12	371575	del	GA	G	G:92 GA:0	CDS	+	3351/3366	1117/1121	frameshift_variant c.3351delA p.Ala1118fs	TETR0B01760	hypothetical protein
12	374765	del	TG	T	T:124 TG:0	CDS	-	58/2388	20/795	frameshift_variant c.58delC p.His20fs	TETR0B01780	hypothetical protein
12	644064	ins	A	AT	AT:98 A:0	CDS	-	687/699	229/232	frameshift_variant c.687dupA p.Tyr230fs	TETR0B02920	hypothetical protein
12	822868	snp	A	G	G:46 A:0	CDS	+	1173/1434	391/477	synonymous_variant c.1173A>G p.Leu391Leu	TETR0B03730	hypothetical protein
12	823019	snp	A	G	G:38 A:0	CDS	+	1324/1434	442/477	missense_variant c.1324A>G p.Thr442Ala	TETR0B03730	hypothetical protein
12	824645	del	CT	C	C:40 CT:0	CDS	+	338/399	113/132	frameshift_variant c.338delT p.Leu113fs	TETR0B03740	hypothetical protein
12	835591	snp	A	G	G:44 A:0	CDS	+	762/942	254/313	synonymous_variant c.762A>G p.Leu254Leu	TETR0B03780	hypothetical protein
12	835702	complex	AGA	G	G:46 AGA:0	CDS	+	873/942	291/313	frameshift_variant&synonymous_variant c.873_875delAGAinsG p.Glu292fs	TETR0B03780	hypothetical protein
12	840158	del	AC	A	A:46 AC:0	CDS	-	561/645	187/214	frameshift_variant c.561delG p.Cys188fs	TETR0B03800	hypothetical protein
12	841416	snp	A	T	T:34 A:0	CDS	-	764/897	255/298	missense_variant c.764T>A p.Ile255Lys	TETR0B03810	hypothetical protein
13	47854	del	AT	A	A:34 AT:0	CDS	+	930/993	310/330	frameshift_variant c.930delT p.Phe310fs	TETR0E00320	hypothetical protein
13	320774	del	AC	A	A:28 AC:0	CDS	-	2235/2238	745/745	frameshift_variant c.2235delG p.Ter746fs	TETR0E01670	hypothetical protein
14	51012	del	AC	A	A:56 AC:0	CDS	+	351/372	117/123	frameshift_variant c.351delC p.Ile118fs	TETR0H00300	hypothetical protein
14	93111	snp	T	A	A:44 T:0	CDS	+	421/714	141/237	missense_variant c.421T>A p.Phe141Ile	TETR0H00450	hypothetical protein
14	108617	snp	A	G	G:38 A:0	CDS	+	838/1548	280/515	missense_variant c.838A>G p.Lys280Glu	TETR0H00510	hypothetical protein
14	721439	del	GT	G	G:88 GT:0	CDS	-	100/2190	34/729	frameshift_variant c.100delA p.Thr34fs	TETR0H03630	hypothetical protein
14	811774	del	GA	G	G:20 GA:0	CDS	-	38/747	13/248	frameshift_variant c.38delT p.Phe13fs	TETR0H04110	hypothetical protein

Supp. Table 1.6 cont. SNP details from snippy of strain 135 compared to Ethanol Red

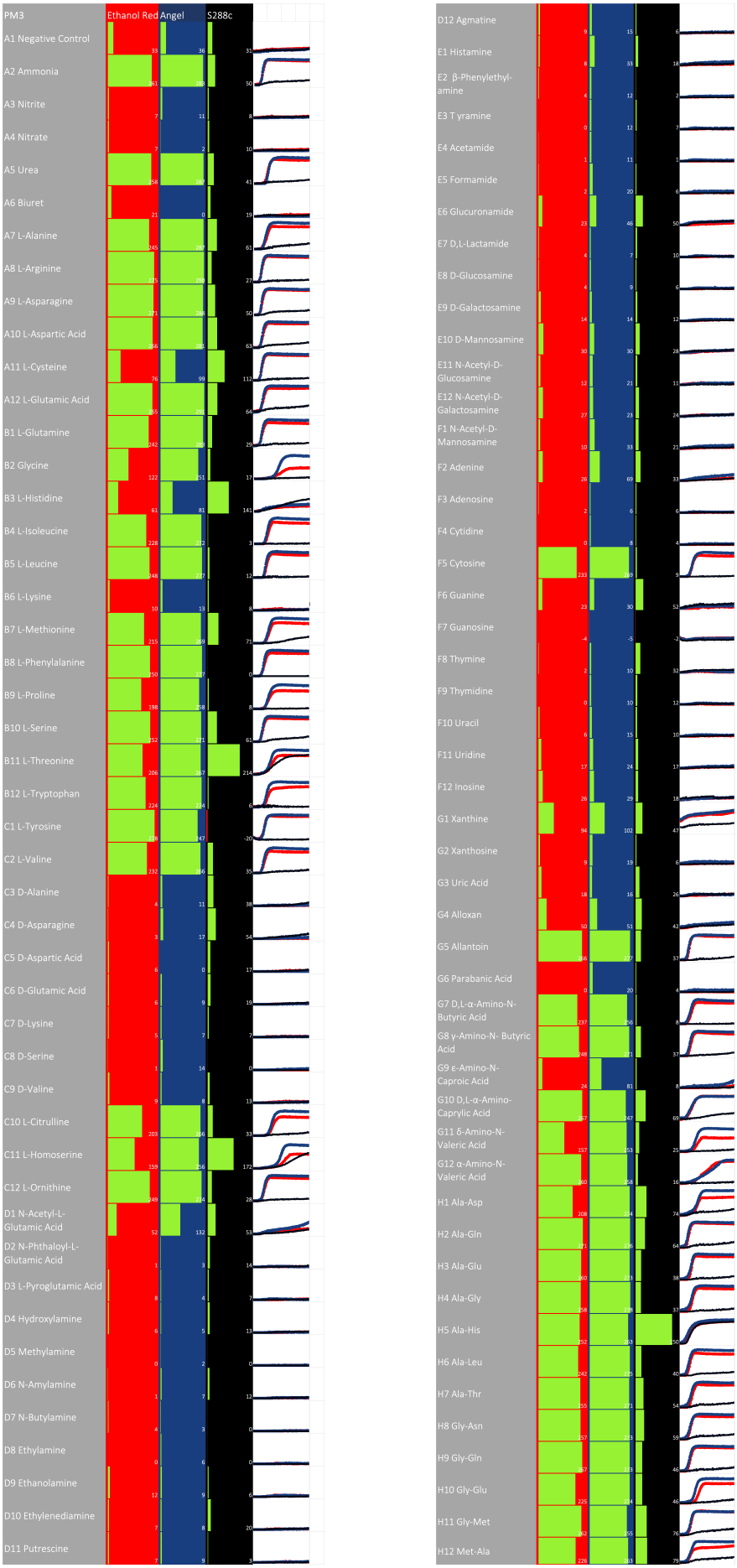
Appendix 2



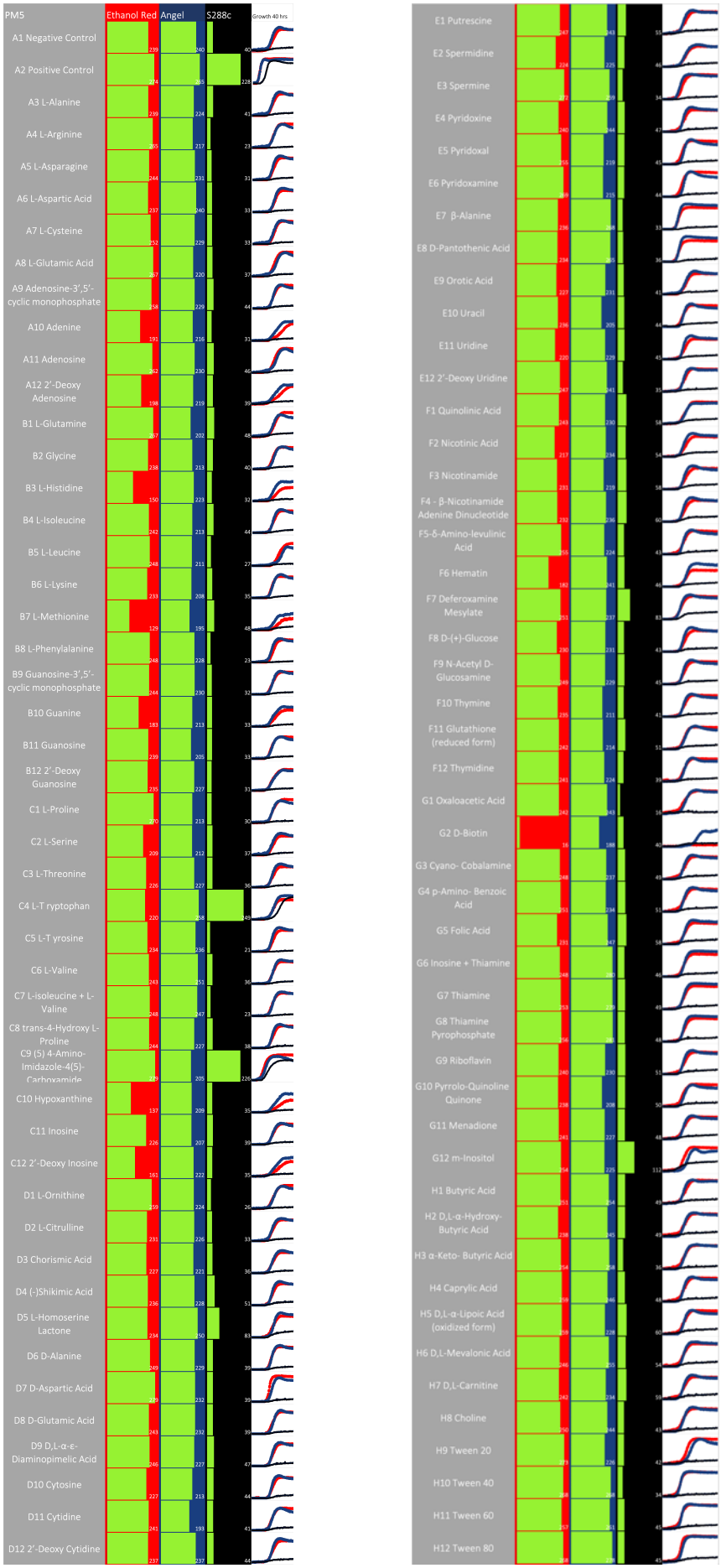
Supp. Figure 2.1Biolog Data for PM1. The final value for endpoint growth is given for three strains on each condition as well as a graph showing each data point in the growth



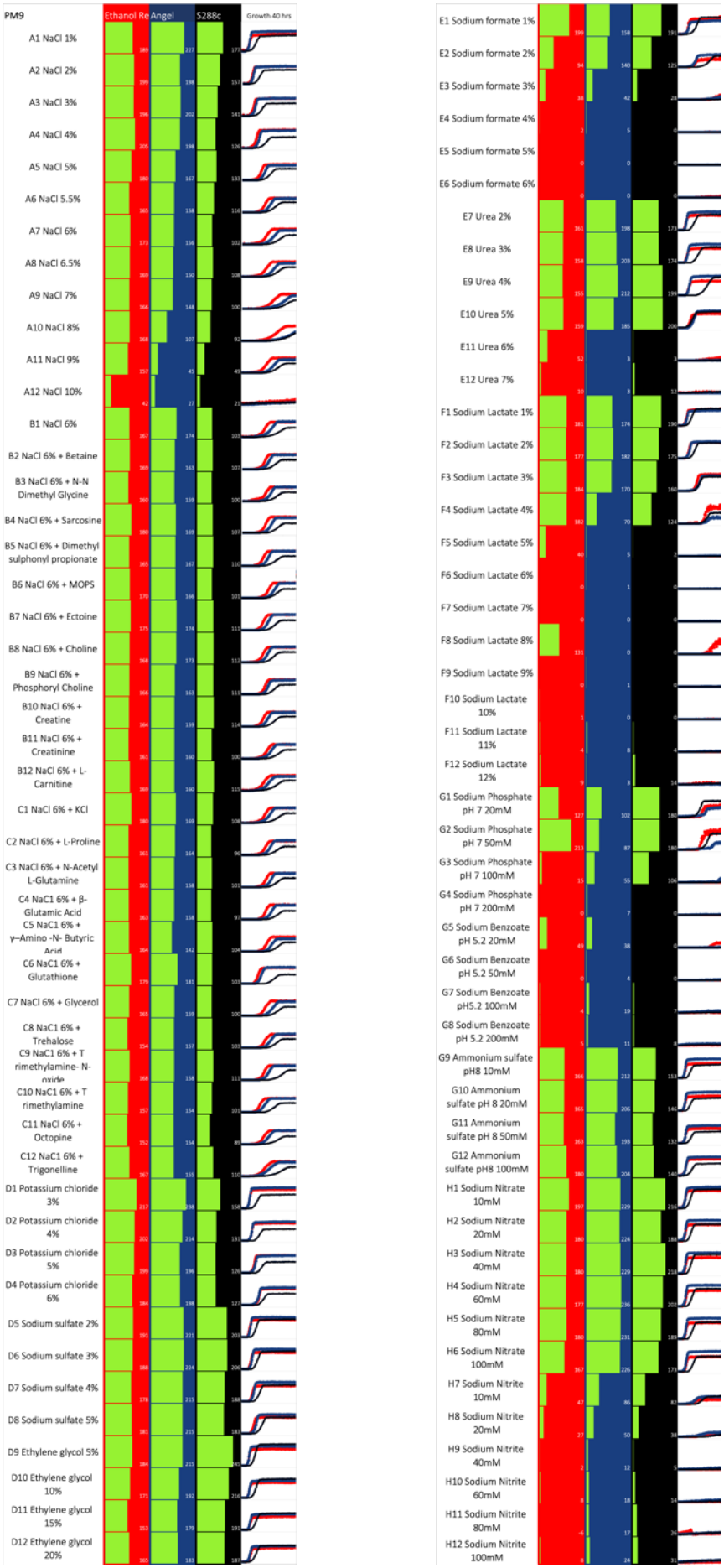
Supp. Figure 2.2 Biolog Data for PM2. The final value for endpoint growth is given for three strains on each condition as well as a graph showing each data point in the growth



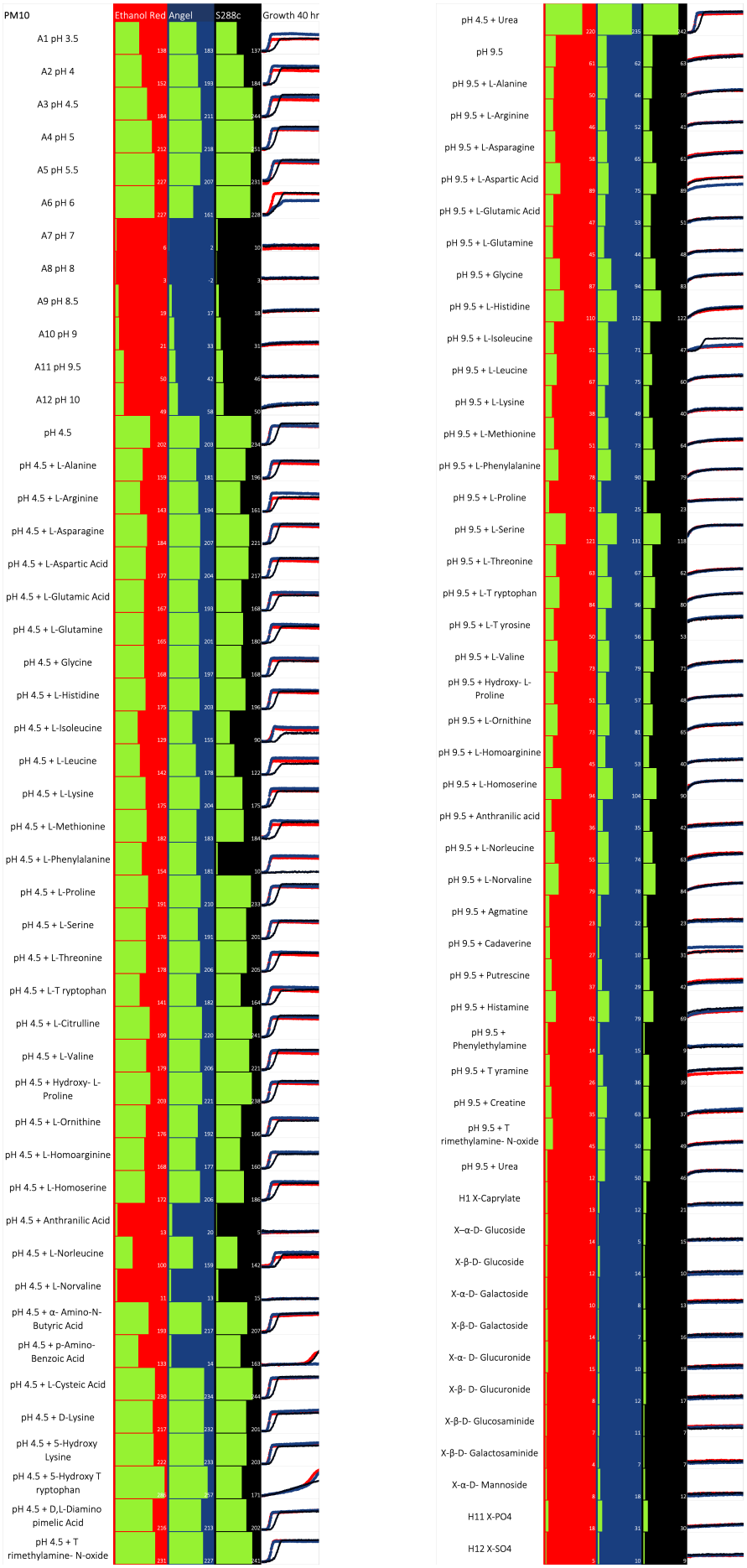
Supp. Figure 2.3 Biolog Data for PM3. The final value for endpoint growth is given for three strains on each condition as well as a graph showing each data point in the growth



Supp. Figure 2.4 Biolog Data for PM5. The final value for endpoint growth is given for three strains on each condition as well as a graph showing each data point in the growth



Supp. Figure 2.5 Biolog Data for PM9. The final value for endpoint growth is given for three strains on each condition as well as a graph showing each data point in the growth



Supp. Figure 2.6 Biolog Data for PM10. The final value for endpoint growth is given for three strains on each condition as well as a graph showing each data point in the growth

Appendix 3.

Introduction to Appendix 3

During this work an unaffiliated body of work was undertaken outside of the collaboration with Ingenza. Due to the similarities in methods and some hypothesis to this work the resulting paper has been included.

Pseudomonas beatica, was initially isolated from diseased wedge sole fish in in Huelva, Spain. Based on 16S and housekeeping gene sequencing and phylogeny as well as phenotypic tests it was proposed by López et al that this was a novel species which they termed *Pseudomonas baetica*. This strain was obtained and sequenced using Ion Torrent sequencing. The resulting data was compiled and taken to the 16th International Conference on Pseudomonas in Liverpool, UK. Here the community was asked to partake in a collaborative project to characterise this strain using the genomic data provided as well as subsequent phenotypic data obtained. My role in this, in addition to performing short read sequencing and initial assembly and a portion of the phenotypic experiments was to organise the data provided from the 28 researchers involved into a manuscript. The method of short read sequencing used here, Ion Torrent, was not used during the PhD project itself, however, the principles of library preparation and techniques involved in both short read and combined long and short-read genome assembly were heavily utilised following this project.

FEMS Microbiology Letters, 365, 2018, fny069

doi: 10.1093/femsle/fny069
Advance Access Publication Date: 22 March 2018
Research Letter

RESEARCH LETTER – Food Microbiology

Community-led comparative genomic and phenotypic analysis of the aquaculture pathogen *Pseudomonas baetica* a390T sequenced by Ion semiconductor and Nanopore technologies

Ainsley Beaton[1,†], Cédric Lood[2,3,†], Edward Cunningham-Oakes[4,†], Alison MacFadyen[5,†], Alex J. Mullins[4,†], Walid El Bestawy[1], João Botelho[6], Sylvie Chevalier[7], Shannon Coleman[8,‡], Chloe Dalzell[1], Stephen K. Dolan[9], Alberto Faccenda[1], Maarten G. K. Ghequire[2], Steven Higgins[10], Alexander Kutschera[11], Jordan Murray[1], Martha Redway[1], Talal Salih[1], Ana C. da Silva[12,#], Brian A. Smith[13], Nathan Smits[3], Ryan Thomson[1], Stuart Woodcock[14], Martin Welch[9,§], Pierre Cornelis[7], Rob Lavigne[3], Vera van Noort[2] and Nicholas P. Tucker[1,*,¶]

[1]Strathclyde Institute of Pharmacy and Biomedical Science, University of Strathclyde, 161 Cathedral Street, Glasgow, G4 0RE, UK, [2]Centre of Microbial and Plant Genetics, KU Leuven, Kasteelpark Arenberg 20, bus 2460, Leuven B-3001, Belgium, [3]Laboratory of Gene Technology, KU Leuven, Kasteelpark Arenberg 20, bus 2460, Leuven B-3001, Belgium, [4]Cardiff School of Biosciences, Cardiff University, Sir Martin Evans Building, Park Place, Cardiff CF10 3AX, UK, [5]Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, Scotland, UK, [6]UCIBIO/REQUIMTE, Laboratório de Microbiologia, Faculdade de Farmácia, Universidade do Porto, Rua de Jorge Viterbo Ferreira no. 228 Porto 4050-313, Portugal, [7]Laboratoire Microbiologie Signaux et Microenvironnement (LMSM), Université de Rouen, 55, rue St Germain, Evreux 27000, France, [8]Lower Mall Research Station, University of British Columbia, 2259 Lower Mall, Vancouver, BC V6T 1Z4, Canada, [9]Department of Biochemistry, University of Cambridge, Hopkins Building, Tennis Court Road, Cambridge CB2 1QW, UK, [10]Department of Plant and Microbial Biology, University of Zürich, Zürich 8008, Switzerland, [11]Department of Phytopathology, Center of Life and Food Sciences, Technical University of Munich, Weihenstephan D-85354, Germany, [12]Centre for Biomolecular Sciences, School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, UK, [13]School of Plant Sciences, The University of Arizona, P.O. Box 210036, Forbes Building, 303 Tucson, Arizona 85721-0036, USA and [14]Department of Biological Chemistry, John Innes Centre, Colney Lane, Norwich NR4 7UH, UK

*Corresponding author: Strathclyde Institute of Pharmacy and Biomedical Science, University of Strathclyde, 161 Cathedral Street, Glasgow G4 0RE, UK. Tel: 0141-548-2861; E-mail: nick.tucker@strath.ac.uk

Received: 5 January 2018; **Accepted:** 21 March 2018
© FEMS 2018. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

†These authors contributed equally to the sequencing or to intensive whole genome data analysis of *P. baetica*. AB generated semiconductor short reads whilst CL produced Nanopore long reads. ECO and AM performed the phylogenetic analysis whilst AMac performed whole genome BLAST analysis and editing.

One sentence summary: Here, we present the high-quality draft genome of *Pseudomonas baetica*, a bacterial pathogen that has been demonstrated to cause disease in fish cultivated in aquaculture facilities.

Editor: Robert Jackson

‡Shannon Coleman, http://orcid.org/0000-0001-9495-6772
#Ana C. da Silva, http://orcid.org/0000-0002-2967-0986
§Martin Welch, http://orcid.org/0000-0003-3646-1733
¶Nicholas P. Tucker, https://orcid.org/0000-0002-6331-3704

ABSTRACT

Pseudomonas baetica strain a390T is the type strain of this recently described species and here we present its high-contiguity draft genome. To celebrate the 16th International Conference on Pseudomonas, the genome of *P. baetica* strain a390T was sequenced using a unique combination of Ion Torrent semiconductor and Oxford Nanopore methods as part of a collaborative community-led project. The use of high-quality Ion Torrent sequences with long Nanopore reads gave rapid, high-contiguity and -quality, 16-contig genome sequence. Whole genome phylogenetic analysis places *P. baetica* within the *P. koreensis* clade of the *P. fluorescens* group. Comparison of the main genomic features of *P. baetica* with a variety of other *Pseudomonas* spp. suggests that it is a highly adaptable organism, typical of the genus. This strain was originally isolated from the liver of a diseased wedge sole fish, and genotypic and phenotypic analyses show that it is tolerant to osmotic stress and to oxytetracycline.

Keywords: *Pseudomonas*; aquaculture; whole genome sequencing; Oxford Nanopore MinIon; Ion Torrent; *Pseudomonas baetica*

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

INTRODUCTION

In September 2017, the biannual conference on the biology of *Pseudomonas* was held in Liverpool, UK. All aspects of *Pseudomonas* biology were reported at the meeting from clinical to environmental microbiology. The vast majority of the *Pseudomonas* literature focuses on just a few organisms, most notably *Pseudomonas aeruginosa, P. fluorescens, P. putida* and *P. syringae,* although the genus is far more diverse than this. In order to contribute to addressing this imbalance, the genome of *P. baetica*, a recently described member of the genus, was sequenced for the Pseudomonas 2017 genomics forum. The resulting genome sequence has been analysed by members of the Pseudomonas 2017 community from across the globe using Slack as a collaboration tool (Perkel 2016).

Pseudomonas baetica strain a390 is one of five Gram-negative organisms that were isolated from a fish disease outbreak in an aquaculture facility in Huelva, Spain, in 2006 (López *et al.* 2012). The infected fish were wedge sole (*Dicologlossa cuneata*), a flat fish that is widely consumed in Andalucía and parts of France where there is increasing interest in sustainable aquaculture of this species. Aquaculture not only provides a sustainable source of protein, but it is also an economically important industry around the world. European commission data reveals that 1.25 million tonnes of food are produced by aquaculture in the EU every year and that over 85 000 people are directly employed in the industry. A disadvantage to aquaculture systems is that the unusually high population density of the farmed fish leads to an elevated risk of infectious disease outbreaks. For this reason, antibiotics such as oxytetracycline (OT) are routinely used as growth promoters in aquaculture fish food pellets (Leal *et al.* 2017). Other members of the *Pseudomonadaceae* are well known to cause disease in aquaculture scenarios, most notably *P. anguilliseptica* is a pathogen of both farmed and wild eels (Joh *et al.* 2013).

Pseudomonas baetica was formally described in 2012 using a combination of 16S rRNA sequencing and phenotypic analyses, allowing a biochemical profile to be generated (López *et al.* 2012). This analysis revealed that *P. baetica* is a member of the *P. fluorescens* group and is particularly closely related to *P. koreensis*. Although infected wedge sole showed no obvious visual signs of *P. baetica* infection, intraperitoneal injection and immersion infection assays yielded 100% and 10% mortality rates, respectively (López *et al.* 2012). Subsequent infection assays with a variety of other fish species demonstrated that *P. baetica* caused higher mortality in wedge sole than other species including sea bass and sea bream, and that there is a temperature-dependent effect on virulence (López *et al.* 2017). In order to rapidly identify *P. baetica* contamination in an aquaculture setting, López *et al.* (2017) have developed immunological and molecular detection assays.

Although *P. baetica* was originally described from an aquaculture infection scenario, a number of recent studies have identified this species and close relatives such as the newly described *P. helmanticensis* and *P. granadensis* in diverse environments in the rhizosphere including bean roots in Iran, and forest soils in Spain (Ramírez-Bahena *et al.* 2014; Pascual *et al.* 2015; Keshavarz-Tohid *et al.* 2017). This suggests that *P. baetica* is more likely to be an opportunistic pathogen of fish as opposed to a genuine marine pathogen. We hypothesised that the draft genome sequence of *P. baetica* would reveal the genetic basis for the traits required to cause disease in wedge sole. The aim of this study is to provide genomic insights into the biology of *P. baetica* and to use this genome as a reflection on the diverse interests of the *Pseudomonas* community.

MATERIALS AND METHODS

Bacterial strains and genome sequence accessions

Pseudomonas baetica a390T originally isolated from the liver of wedge sole fish, *D. cuneate*, and described by Lopez *et al.* (2012), was obtained from DSMZ (DSM no. 26532) and is considered to be the type strain. Unless stated otherwise, all phenotypic

comparisons were performed against *P. fluorescens* PF01, *P. putida* KT2440 and *P. aeruginosa* PA14. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PKLC00000000. The version described in this paper is version PKLC01000000. Raw sequencing reads were also deposited and can be found with accession numbers SRR6792524 and SRR6792523 for the Nanopore and Ion Torrent reads, respectively.

DNA extraction, library preparation and Ion Torrent PGM Sequencing

DNA extraction was carried out using ISOLATE II genomic DNA kit (BIOLINE) using the standard protocol for bacterial cells. Fragmentation was carried out using the standard protocol from NEBNext Fast DNA Fragmentation & Library Prep Set for Ion Torrent and adapter-ligated DNA was prepared using barcode set 6. Size selection was carried out using the E-gel ibase (Invitrogen), and analysed using the 2100 Bioanalyzer (Agilent Genomics). A fragment size of 424 bp was carried forward. Template amplification and enrichment were carried out using the standard Ion Torrent protocol using the Ion OneTouch 2 system (ThermoFischer). The standard sequencing protocol for Ion Torrent was used and DNA library was loaded onto an Ion316 chip v2 with an ISP loading percentage of 87%.

Nanopore sequencing

Long reads by Nanopore sequencing were obtained using a *MinION* sequencer from Oxford Nanopore Technology (ONT) with the goal to improve the genome assembly of *P. baetica*. The strain was grown in LB medium and grown at 30°C to an OD_{600} of 0.850 and its genomic DNA was isolated using the Mo Bio DNAeasy UltraClean Microbial kit. Quality of the gDNA was checked using a NanoDrop and an agarose gel for integrity. The gDNA was sheared mechanically to an average fragment size of 10 kbp using a Covaris *gTube*. The sequencing library was prepared using the 1D ligation protocol from ONT with native barcoding of the sample. The result was sequenced on a R9.4 flowcell for a period of 8 h.

Combined Genome Assembly of Ion PGM and Nanopore Reads

The Ion PGM reads were checked for quality using *FastQC* and processed with *BBDuk* in order to remove potential adapter contamination, size exclusion (50> read length <250), and for trimming (phred score < 28) (Andrews 2014). Basecalling of the Nanopore reads was performed with the software *Albacore* v2.1.3 from ONT, and the barcodes were removed using *Porechop*. Read length distribution and quality of the Nanopore reads was assessed using *NanoPlot*.

The genome was first assembled with *SPAdes*, version 3.11.1 using default parameters (Bankevich *et al.* 2012) using Ion PGM reads only, then with *Unicycler*, version v0.4.3 (Wick *et al.* 2017) also using default parameters with both Ion PGM and Nanopore reads. For the latter, the genome was polished with Racon (Vaser *et al.* 2017) and Pilon (Walker *et al.* 2014). The quality of both assemblies was assessed with *QUAST* (Gurevich *et al.* 2013).

Genomic island prediction and genome visualisation

The Ion PGM and the hybrid assemblies were annotated with *Prokka* (Seemann 2014) using a custom protein database built from the *Pseudomonas* genus. The annotations were then searched for prophages using *PHASTER* (Arndt *et al.* 2016), and genomic islands using the *IslandViewer* and antiSMASH (Bertelli *et al.* 2017; Blin *et al.* 2017). For the latter, the reference genome used for the alignment was that of *P. koreensis* strain D26 as it was deemed phylogenetically close to *P. baetica*. CCT maps were created using CCT version 1.0 using default settings with the exception of the following commands; -t –cct –custom scale_blast=F _cct_blast_thickness=60 backboneThickness=10 tickThickness=15 rulerFontSize=100 featureThickness=150 tickLength=30 legend=F details=F. A features file was used for the 'beige blocks', and a 'labels_to_show' file was used to label the CCT map (Grant, Arantes and Stothard 2012).

Phylogenetic analysis

Trees were assembled using Mash v1.1.1 (Ondov *et al.* 2016) fast genome and metagenome distance estimation using the MinHash algorithm, and visualised in FigTree v1.4.3 (Rambaut 2012, also available via GitHub). Reference genomes were obtained from NCBI. Average nucleotide identity (ANI) was performed using PyANI (Gupta *et al.* 2016). Data were processed using CLIMB (Connor *et al.* 2016).

Azocasein assay

Overnight cultures of *P. aeruginosa* UCBPP-PA14, *P. baetica* a390T, *P. fluorescens* Pf0-1 and *P. putida* KT2440 were centrifuged at 12 000× *g* for 5 min and 100 μl of supernatant was pipetted into centrifuge tubes containing 400 μl of reaction mixture; 200 μl of 2% azocasein in 0.5% bicarbonate sodium buffer and 200 μl of 0.5% sodium bicarbonate buffer. The reaction mixture was then incubated at 30°C for 10, 70 and 140 min.

Following incubation, 0.5 ml of 10% trichloroacetic acid was added to terminate the reaction. Tubes were then vortexed for 5 min followed by centrifuging for 5 min at 4°C and 10 000× *g*. Five hundred microlitre of supernatant was then added to cuvettes containing 500 μl of 0.5 M NaOH, and absorbance was measured at 440 nm in an Ultrospec 3100pro spectrophotometer. All samples were run in triplicate.

Skim milk assay

Five microlitre of overnight cultures of *P. aeruginosa* UCBPP-PA14, *P. baetica* a390T, *P. fluorescens* Pf0-1 and *P. putida* KT2440 were inoculated on to MOPS minimal media agar plates, supplemented with dehydrated skimmed milk and incubated for 72 h at 30°C. Photographs were taken at 24-h time intervals. Iron supplement, $FeCl_2$, was not included in the 10× stock, and 5g/L of agar and skim milk were added in order to observe clear halo formation.

OT resistance assay

Pseudomonas baetica was tested for antibiotic sensitivity by agar dilution according to the European Committee for Antimicrobial Susceptibility Testing guidelines (European Committee for Antimicrobial Susceptibility Testing (EUCAST) of the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) 2000). Resistance to OT was tested due to its frequent use as a growth promoter in aquaculture (Leal *et al.* 2017). Mueller-Hinton agar plates with final concentrations ranging from 0.125 to 128 μg mL^{-1} were inoculated with *P. baetica* a390T suspensions diluted in 0.85% NaCl to a turbidity equivalent of a 0.5 McFarland standard. Plates were incubated overnight at 30°C. *Pseudomonas aeruginosa* PAO1 was used as a positive control.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

Table 1. Assessment of genome assembly quality for both approaches reveals a vast improvement in the assembly despite a low nanopore sequencing coverage (9X).

Metric	Ion PGM only	Hybrid
Assembly length	6602 908	6773 804
No. of contigs (>1000 bp)	338	16
N50	36 090	973 739
N75	19 947	745 006
Longest contig	122 046	1654 292

RESULTS

Sequencing results and assembly

The initial Ion Torrent sequencing produced over 3.5 million barcoded reads with a mean read length of 299 bp. SPAdes assembly resulted in final contig number of 338. The Nanopore sequencing run yielded around 22 000 reads, of which 7771 passed quality control during the basecalling process. The average phred quality score of the reads is 11 and their length distribution was around 9 kbp (see Figure S1, Supporting Information). With a genome size of about 6.6 Mbp, estimated from an initial Ion PGM assembly, we infer a genome coverage of about 9× for the Nanopore sequencing data (60 Mbp), and 80× with Ion PGM data (540 Mbp after QC).

The hybrid approach that combines both the dataset of short Ion PGM reads with the long Nanopore reads significantly improved the genome assembly over that using Ion PGM reads only giving 16 contigs over 1000 bp (Table 1). Quast analysis of the combined assembly gives an N50 value of 973739 which is higher than 96% of all draft genomes in The Pseudomonas Genome database.

The Ion PGM-only and the hybrid assembly graphs were also inspected visually using *Bandage* (Wick *et al.* 2015) in order to assess the fragmentation of the genome assemblies and the genomic organisation, including the presence of potential plasmids (see Figures S2 and S3, Supporting Information). After gene annotation with *Prokka*, plasmid-associated genes like recombinases, transposases and integrases were searched for and were not present or were not indicative of being plasmid associated. The high continuity assembly provided by the combinatorial approach of Ion Torrent and Nanopore technologies allowed for a comprehensive analysis of the genome.

Phylogenetic placement and ANI of *P. baetica* a390T

We have produced a phylogeny to place the sequenced *P. baetica* a390T genome against reference genomes from 19 species clades, comprising the entire genus *Pseudomonas* (Fig. 1B) (Gomila *et al.* 2015). This was done using Mash, a pipeline that converts a genome into a sketch (or groups of k-mers). Sketches are then compared to produce a Jaccard index, which is generated based upon shared k-mers between genomes (Ondov *et al.* 2016). Our phylogeny places *P. baetica* amongst the *P. koreensis* subclade, as previously shown in phylogenies constructed using partial 23S rRNA *P. koreensis* gene sequences. A *P. koreensis* phylogeny constructed in the same manner places *P. baetica* closest to *Pseudomonas sp.* Irchel 3E19, a strain isolated from a pond in Zurich, Switzerland, suggesting that this isolate is in fact a strain of *P. baetica* (Butaitė *et al.* 2017). The next most closely related organism is *P. koreensis strain* CI12, which was recently co-isolated with *Bacillus cereus* from a soybean plant rhizosphere (Bravo, Lozano and Handelsman 2017). ANI was used to compare the whole genome sequence of *P. baetica* against the six available reference genomes for *P. koreensis* as well as strain Irchel 3E19, and confirms that it is most closely related to Irchel 3E19 and CI12, with an ANI of 98.02% and 89.6% similarity, respectively (Fig. 1A) (Arahal 2014). We also used JSpeciesWS as an independent ANI method and this analysis agreed with those described above (see supplementary ANI tables) (Richter *et al.* 2016). Finally, BLAST was used to compare the *gyrB* and *rpoD* genes sequences of *P. baetica* a390T and CI12 against available partial *rpoD* and *gyrB* sequences available for *P. baetica* (López *et al.* 2012). This showed a 93.6% similarity in *rpoD* and 95.0% similarity in *gyrB* between CI12 and the whole genome-sequenced *P. baetica* a390T.

Genomic islands, biosynthetic gene clusters and prophages detected in the *P. baetica* genome

Genomic islands can be readily detected when long contiguous regions are assembled. As such, searching for prophages in the highly fragmented, ion PGM-only assembly did not yield any results, whereas inspection of the hybrid assembly showed four intact prophages in the genome of *P. baetica*, as well as five incomplete ones (Table 2a). Two integrative conjugative elements were also detected. A combination of IslandViewer and antiSMASH were also used to predict genomic islands and the secondary metabolite biosynthetic potential of *P. baetica* (Bertelli *et al.* 2017; Blin *et al.* 2017). These results are summarised on the CG View Comparison Tool map in Fig. 2, and a more detailed summary of the antiSMASH results is provided in Table 2b (Grant, Arantes and Stothard 2012). Many of the loci referred to throughout this manuscript are labelled around the CCT map in Fig. 2 as well as the results of IslandViewer. The CCT BLAST analysis confirms the close phylogenetic relationship between *P. baetica* a390 and *Irchel* 3E19 (Fig. 2). Of the nine biosynthetic gene clusters (BGCs) predicted by antiSMASH, four are Nrps (non-ribosomal peptide synthetases) biosynthetic gene clusters. Clusters 3 and 8 had relatively low similarities to the nearest known clusters; however, cluster 6 shared a high similarity with an orfamide BGC with 80% of genes in this cluster sharing homology with the putisolvin biosurfactant BGC. Putisolvin has been implicated in motility (Cárcamo-Oyarce *et al.* 2015) and our preliminary experiments have demonstrated that *P. baetica* is capable of both swimming and swarming (see Figure S4, Supporting Information). Since putisolvin biosynthesis in certain strains of *P. putida* is known to be a quorum sensing (QS)-regulated phenotype, we searched for homologues of known QS systems in the *P. baetica* genome (Dubern *et al.* 2008). No convincing homologues of the LasR or RhlR regulated systems from *P. aeruginosa*, or the PpuI, PpuR, PpuA, RsaL system from *P. putida* were found.

Bacteriocins, tailocins and modular bacteriocins of *P. baetica*

Pseudomonads are capable of producing a variety of antagonism-mediating peptides and proteins. A subset of these molecules are bacteriocins, and assist in the elimination of phylogenetically related competitors (Ghequire and de Mot 2014). Historically, these compounds have been mainly studied in *P. aeruginosa*, there termed pyocins (Michel-Briand and Baysse 2002), but they have been equally detected and characterized in a number of other *Pseudomonas* species. Evolutionarily related to bacteriophage tails, tailocins are functional standalone units—rather than defective prophages—lacking an

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021



Figure 1. Placement of novel *baetica* genome in the genus *Pseudomonas* based upon ANI, and shared k-mers A) ANI heatmap generated the Python3 module pyani. The sequenced baetica genome, Pseudomonas_sp._Irchel_3E19 and reference genomes from the closely related *koreensis* subclade were subject to ANI analysis. B) A Mash-based tree generated from reference genomes ($n = 86$) from 19 species clades, comprising the entire genus *Pseudomonas*. This tree was generated based upon the Jaccard index, calculated from shared k-mers. *Cellvibrio japonicum* Ueda107 was used as outgroup.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

Table 2a. List of prophage regions detected in the hybrid assembly.

Contig no.	Length	Completeness	Score	No. of proteins	Region position	Most common phage	GC %
1	6.7 Kbp	Questionable	70	6	681 684–688 458	PHAGE_Stx2_c_1717_NC_011357(3)	53.77%
2	6.5 Kbp	Incomplete	40	10	512 098–518 615	PHAGE_Entero_phi92_NC_023693(2)	60.11%
2	10.2 Kbp	Incomplete	30	8	742 363–752 628	PHAGE_Clostr_phiCT453A_NC_028991(3)	56.63%
2	5.5 Kbp	Incomplete	60	6	925 750–931 322	PHAGE_Stx2_c_1717_NC_011357(3)	54.15%
3	40.7 Kbp	Intact	150	48	383 269–424 017	PHAGE_Vibrio_VP882_NC_009016(12)	58.74%
3	57.2 Kbp	Intact	150	49	430 072–487 323	PHAGE_Vibrio_vB_VpaM_MAR_NC_019722(8)	58.69%
4	35.1 Kbp	Intact	150	42	357 957–393 153	PHAGE_Pseudo_phiCTX_NC_003278(18)	57.32%
4	50.1 Kbp	Intact	150	70	564 492–614 686	PHAGE_Entero_Arya_NC_031048(10)	56.98%
5	21.3 Kbp	Incomplete	40	11	26 364–47 679	PHAGE_Pseudo_vB_PsyM_KIL1_NC_030934(6)	49.79%
5	8.7 Kbp	Incomplete	50	13	540 790–549 505	PHAGE_Bacill_BMBtpLA_NC_028748(1)	57.37%

accompanying phage head (Ghequire and de Mot 2015). *Pseudomonas* phage tail-like bacteriocins are typically recruited to *trpE-trpG* and *mutS-cinA* intergenic regions (Ghequire and de Mot 2015). Inspection of the corresponding regions in *P. baetica* reveals the presence of two adjacent R-type gene clusters at the *mutS-cinA* locus. Phylogenetic assessment and gene synteny is indicative of an Rp3-Rp4 organisation, and a similar tailocin configuration was recently described in *P. chlororaphis* 30-84 (Dorosky *et al.* 2017). An interesting observation is that four putative tail fibres are likely present in the Rp4 cassette of *P. baetica*, whereas usually only one or two of these can be detected in tailocins.

Referred to as S-type pyocins in *P. aeruginosa*, modular bacteriocins represent a heterogeneous group of polymorphic toxins (Zhang *et al.* 2012; Jamet and Nassif 2015), comprising a receptor-binding domain, a segment enabling membrane transfer and a toxin domain at the carboxy-terminus (Ghequire and de Mot 2014). To protect from self-inhibition, bacteriocin producer strains co-express an immunity gene, typically encoded adjacent to the bacteriocin killer gene. *Pseudomonas* nuclease bacteriocins (DNase, tRNase, rRNase) are consistently equipped with a central pyocin_S domain (Pfam06958), and suggested to play a pivotal role in translocation of the toxin to the cytoplasm (Ghequire and de Mot 2014; Sharp *et al.* 2017). Searching for this domain revealed an array of putative modular bacteriocins in *P. baetica*, seven of which host an HNH-DNase and three an rRNase domain. In one bacteriocin, a putative (non-HNH-type) DNase homologous to the killer domain found in carocin D—but distinct from the non-HNH-type in pyocin S3—was identified.

Two other predicted *P. baetica* bacteriocins with homology to a toxin module of a contact-dependent inhibition cassette (type IV) in *Burkholderia pseudomallei* were also detected. No bacteriocins acting at the periplasmic level (colicin M-like, Pfam14859; pesticins, Pfam16754; or pore formers (ColIa/ColN), Pfam01024) could be retrieved.

Lipid A and core oligosaccharide biosynthesis gene clusters

Putative homologues of the confirmed lipid A biosynthesis (Raetz pathway) enzymes in *P. aeruginosa* PAO1 could be identified in the sequenced genome. Comparison to different *Pseudomonas* proteomes revealed sequence identities of the putative proteins ranging from 59% to 97%. Those high-sequence identities and the hydrocarbon ruler motive in putative LpxA (UDP-N-acetylglucosamine acyltransferase) hint towards a lipid A with a C10 acyl chain at position 3 and 3′ (Smith *et al.* 2015). Besides, C12 secondary acyl chains or C12 primary acyl chains at position 2 and 2′ are possible due to sequence homology of putative LpxD and the presents of possible HtrB1 and HtrB2 homologs (Lam *et al.* 2011; Hittle *et al.* 2015). In general, it is likely that *P. baetica* possesses a lipid A structure similar to known structures from *P. aeruginosa* when possible post-synthesis modifications are not taken into account (Knirel *et al.* 2006).

Further genome analysis resulted in the identification of a gene cluster homologous to the core oligosaccharide gene cluster which was described and studied in *P. aeruginosa*. Direct comparison of the cluster revealed that *P. baetica* lacks homologues of the *P. aeruginosa* genes PA4998 (putative kinase) and PA5008 (heptose kinase), which possibly suggests a less phosphorylated core oligosaccharide (Lam *et al.* 2011). Reciprocal BLAST searching of different *Pseudomonas* proteomes resulted in sequence identities between the wide range of 14% and 98%. The highest overall sequence similarity was observed to *P. fluorescens* Pf0-1 core oligosaccharide biosynthesis proteins, ranging from 85% to 98% which is consistent with the phylogeny. In general, the cluster structure and presence of putative glycosyltransferases (WapR, WapH, WaaG, WaaC, WaaF) indicates the presence of an inner core oligosaccharide structure similar to *P. aeruginosa* (Knirel *et al.* 2006).

Extracellular protease and effector secretion by *P. baetica*

Extracellular protease secretion is an important virulence determinant of *P. aeruginosa*. In order to determine if *P. baetica* has a similar phenotype, azocasein and skimmed milk protease assays were performed with *P. aeruginosa* UCBPP-PA14, *P. baetica*, *P. fluorescens* Pf01, and *P. putida* KT2440. As expected, *P. aeruginosa* UCBPP-PA14 secreted prodigious amounts of protease (see Figures 5 and 6, Supporting Information). *Pseudomonas baetica* a390T secreted significantly less protease than *P. aeruginosa* but marginally more than the other strains. This phenotype was most evident on MOPS minimal milk agar after 72 h (see Figure S6b, Supporting Information).

A major virulence determinant in Gram-negative pathogens is the ability to inject effector proteins into host cells using a type III secretion system (T3SS) (Hauser 2009). BLAST analysis using the *P. aeruginosa* Psc T3SS proteins revealed a T3SS to be present in the *P. baetica* genome, providing a possible explanation for the virulence of this strain as hypothesised above.

OT resistance in *P. baetica*

Following overnight incubation, antibiotic plates were examined for bacterial growth, and the minimum inhibitory concentra-

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

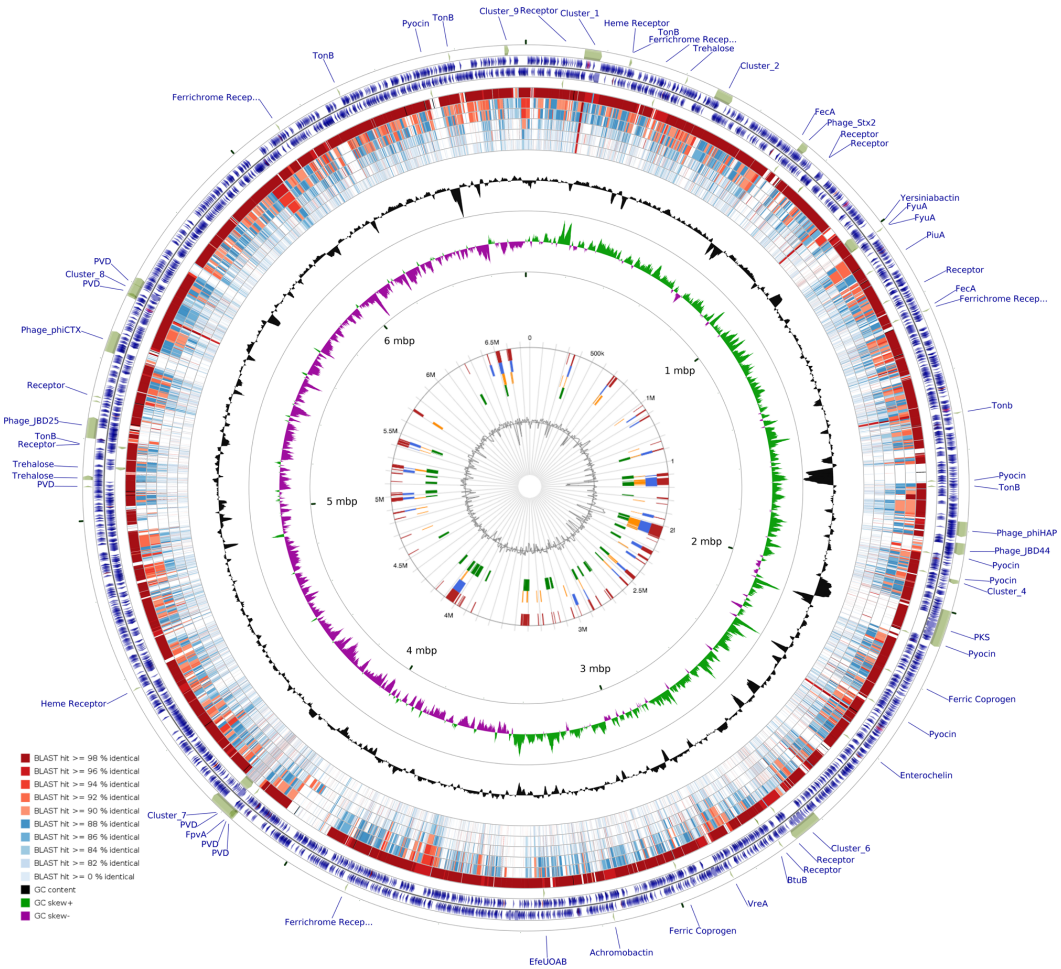


Figure 2. Overview of the genomic architecture of *P. baetica* a390T. CGView Comparison Tool was used to plot the percentage sequence identity of closely related strains. The outer labels and beige blocks refer to loci that are mentioned throughout the paper. The *P. baetica* a390T genome is split into the forward and reverse strands and the coding sequences are represented by blue arrows. The next 6 rings represent the percentage BLAST identity of the genomes of the following closely related strains; from the outside ring to inside the order is *P. sp Irchel* 3E19, *P. koreensis CI12*, *P. koreensis D26*, *P. fluorescens* SBW25, *P. putida* KT2440, *P. aeruginosa UCBPP-PA14*. The percentage identity is indicated by the colour of the 6 BLAST rings as indicated by the key on the bottom left hand side of the figure. Moving inwards, the next two rings indicate the changes in %GC content and the GC skew respectively. Finally, the central diagram indicates the genomic islands predicted by the IslandViewer package where Red = predicted by multiple methods, blue = IslandPath-DIMOB, orange = SIGI-HMM and green = IslandPick.

Table 2b. Secondary metabolite biosynthetic clusters predicted by antiSMASH.

Cluster	Type	Cluster size (bp)	Most similar known biosynthetic gene cluster	Percentage similarity
1	Unknown	43 395	Mangotoxin	71
2	Arylpolyene	43 613	APE_Vf	40
3	Nrps	63 804	Nataxazole	11
4	Bacteriocin/Ripp	10 887	–	–
5	Transatpks-Nrps	94 211	Sorangicin	13
6	Nrps	77 485	Orfamide/Putisolvin	70
7	Terpene-Nrps	73 119	Pyoverdine	16
8	Nrps	52 998	Pyoverdine	22
9	Bacteriocin/Ripp	10 845	–	–

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

tion (MIC) of OT for *P. aeruginosa* PAO1 was determined as 1 μg mL^{-1}, a quarter that of *P. baetica* a390T. OT is a protein synthesis-inhibiting antibiotic belonging to the tetracycline class, and is one of the most commonly used antibiotics in aquaculture (Leal *et al.* 2017).

When *P. baetica* was cultured on media containing OT, it was able to grow in the presence of higher concentrations than *P. aeruginosa* PAO1. The minimum inhibitory concentration of OT required to inhibit the growth of *P. baetica* was determined as 4 μg mL^{-1}. The ability of *P. baetica* to tolerate higher concentrations of OT than a lab strain of *Pseudomonas* may be due to the use of the antibiotic in aquaculture, especially as an additive in fish food. OT resistance (OT^R) has been observed previously in bacterial species colonising pelletised fish food in Chile, and also in half of commercial fish food samples analysed in a 1995 study of OT^R in fish farm sediments (Kerry *et al.* 1995; Miranda *et al.* 2003).

Glycerol-3-phosphate utilisation in *P. baetica*

A notable characteristic which differentiates *P. baetica* from closely related *Pseudomonas* species is an inability to grow on glycerol-3-phosphate (G3P) as a sole carbon source (López *et al.* 2012). Several other pseudomonads, such as *P. fluorescens* and *P. aeruginosa* can use G3P as sole carbon sources. However, *P. baetica* can use glycerol as a sole carbon source. This suggested that *P. baetica* suffers from a specific deficit in G3P uptake, not metabolism. Uptake of G3P is known to occur using the specific MFS transporter GlpT (Fann, Busch and Maloney 2003).

The newly sequenced genome allowed us to examine the *P. baetica* growth deficiency on G3P in greater detail. Remarkably, *P. baetica* appears to harbour a *glpT* homologue, which is quite similar to GlpT from other pseudomonads (see Figure S7, Supporting Information). The structure of GlpT has been elucidated for *Escherichia coli*, allowing us to understand which residues are important for substrate recognition (Huang *et al.* 2003). All residues which have been identified as important for functionality in previous studies are unchanged in *P. baetica*. Mutation of K80 to alanine killed heterologous G3P-Pi transport of the protein reconstituted into proteoliposomes, and mutation of H165 resulted in a transport rate that was only ~6% of that catalysed by wild-type protein (Law *et al.* 2008). These data suggest that the presence of a *glpT* homolog does not necessarily equate to G3P uptake, highlighting the necessity of experimental validation. This result may be due to a lack of induction of the *P. baetica glpT* gene upon exposure to G3P.

The broad-spectrum bactericidal antibiotic, fosfomycin, is taken up actively into bacterial cells via GlpT and the glucose-6-phosphate transporter UhpT, and inhibits the initial step in cell wall synthesis (Kahan *et al.* 1974). Inactivation of *glpT* was shown to confer increased fosfomycin resistance in *P. aeruginosa*, with an apparent lack of fitness cost. These authors isolated 10 independent spontaneous mutants, obtained on LB-fosfomycin, which harboured mutations in the *glpT* gene. These mutants included several deletions, frameshifts and three single amino acid changes, Gly_{137} to Asp, Thr_{336} to Pro and Met_{366} to Ile (Castañeda-García *et al.* 2009). Indeed, mutations in the *Staphylococcus aureus glpT* gene were recently shown to be the major determinant of fosfomycin resistance in this organism (Xu *et al.* 2017). The high fosfomycin resistance and low biological fitness cost resulting from the loss of G3P uptake suggests that *P. baetica* may have an evolutionary advantage in a fosfomycin-rich environment. Future work will be necessary to confirm precisely how G3P uptake has been lost in *P. baetica*.

Orthologs of MexAB-OprM and MexEF-OprN resistance-nodulation cell division efflux systems in *P. baetica*

The MexAB-OprM and MexEF-OprN systems are members of the resistance-nodulation cell division multidrug efflux pump family (Poole 2011). BLASTp was used to compare the amino acid sequences encoded by the *mexEF-oprN* and *mexAB-oprM* operons in *P. aeruginosa* strain PAO1 with homologues in *P. baetica*.

MexAB-OprM is known to play a key role in efflux-mediated resistance to a wide variety of compounds such as antibiotics and solvents in *P. aeruginosa* (Li, Zhang and Poole 1998; Poole 2011). Similarly, the MexEF-OprN efflux substrates include antibiotics and the QS precursor 4-hydroxy-2-heptylquinoline (Lamarche and Déziel 2011; Llanes *et al.* 2011). *Pseudomonas baetica* sequences producing significant alignment to MexAB-OprM and MexEF-OprN were identified with between 69%–79% and 72%–88% identity to *P. aeruginosa* strain PAO1, respectively. Interestingly, the arrangement of the *mexAB-oprM* operon in *P. baetica* bears more similarity to *P. fluorescens*, where the adjacent regulatory gene encodes a member of the TetR regulator family, unlike *P. aeruginosa* which encodes the MarR family regulator MexR in this position. However, it should be noted that *P. baetica* does have a MexR homologue located adjacent to a third efflux operon which has significantly less homology (8%–31%) to *mexAB-oprM*.

Outer membrane porins in *P. baetica*

In addition to the Mex-associated porins, the *P. baetica* genome was searched for genes coding for other outer membrane porins. There are 26 porin genes in *P. aeruginosa* involved in different important biological functions: outer membrane stability (OprF, OprH) and nutrient uptake (OprB, OprP, and the porins of the OprD family), which have recently been reviewed (Chevalier *et al.* 2017). The *P. baetica* genome revealed only 17 porin genes, but the major porins described in *P. aeruginosa* are conserved. The *oprF* gene coding for the major *Pseudomonas* porin is present in *P. baetica* and the genomic context as well: *estX-menG-cmaX-cfrX-cmpX-sigX-oprF* where EstX is an esterase, MenG a RNAse inhibitor, CmaX a CorA-like magnesium transporter, CfrX a putative anti-sigma factor, CmpX a mechano-sensitive channel, and SigX an ECH-sigma factor. OprF has a C-terminal periplasmic domain interacting with peptidoglycan and a barrel N-terminal domain and it is a major contributor to the envelope stability. Other conserved porin genes are *oprG* and *oprH*, which code for small porins with only eight strands. OprG belongs to the OmpW family and has been proposed to be involved in Fe^{2+} diffusion to the periplasm (Catel-Ferreira *et al.* 2016). OprH is induced by low Mg^{2+} and contributes to the outer membrane stability via its interaction with LPS (Kucharska *et al.* 2015). As in *P. aeruginosa*, *P. baetica* also has three genes for the glucose porin OprB, including one in the conserved *gltBFGKoprB* operon (Chevalier *et al.* 2017). Surprisingly, the *P. baetica* genome does not contain genes for the OprO and OprP phosphate uptake porins of *P. aeruginosa*. The OprD family comprises 19 members in *P. aeruginosa* sub-divided into two sub-families, the OccD being involved in the uptake of basic amino acids and the OccK for the uptake of negatively charged cyclic molecules (Eren *et al.* 2012). The comparison with *P. aeruginosa* porins via BLASTP is complicated by the fact that many members of the OprD family show significant relatedness. Nevertheless, the survey of the *P. baetica* genome revealed five members of the OccD and six of the OccK family.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

Osmotic stress resistance of *P. baetica*

Since *P. baetica* a390T was originally isolated from a marine aquaculture scenario, we have investigated its genotypic and phenotypic ability to overcome salt stress. Trehalose is a stress-relieving disaccharide which accumulates within the bacterial cell in response to stress. It has been widely studied for its role as a compatible solute that confers protection against osmotic stress. It is thought that trehalose rehydrates the cell following water loss, preventing the loss of cellular functions (De Smet *et al.* 2000). It has also been reported that trehalose is important for pathogenicity of *Pseudomonas* spp. (Freeman, Chen and Beattie 2010; Djonović *et al.* 2013). Bacteria possess three main pathways for trehalose biosynthesis: OtsAB, TreYZ and TreS.

The OtsAB pathway is the most common route of trehalose biosynthesis in bacteria. OtsA catalyses the synthesis of trehalose 6-phosphate from UDP-glucose or ADP-glucose and glucose 6-phosphate. The phosphate group is then removed by OtsB to give trehalose (De Smet *et al.* 2000). The TreYZ and TreS pathways produce trehalose using different substrates. TreY converts reducing maltosyl units of maltooligosaccharides to give α-1,1 terminal moieties which are then cleaved by TreZ to produce trehalose. The TreS pathway isomerises maltose into trehalose, catalysed by the trehalose synthase TreS, but its utility in vivo is context-dependent (Miah *et al.* 2013).

Comparative genomics have shown that in common with other *Pseudomonas* spp., *P. baetica* lacks the genes for the OtsA/OtsB pathway. However, *P. baetica* possesses the enzymes of the TreYZ and TreS pathways.

Although TreS is known to be capable of generating trehalose in some bacteria, it is predicted that the function of TreS in *P. baetica* is to catabolise trehalose. Phenotypic observations suggest that the ability of *P. baetica* to utilise trehalose as a sole carbon source is dependent on other conditions. In a study by López *et al.*, it was reported that *P. baetica* was unable to utilise trehalose. Because of its importance to osmotic stress, we repeated the Omnilog analysis for trehalose and observed an intermediate phenotype and further analysis on defined media shows that *P. baetica* is capable of utilising trehalose.

The *treS* gene is fused with the maltokinase-encoding *pep2* gene (Chandra, Chater and Bornemann 2011), consistent with a TreS/Pep2 fusion protein converting trehalose into α-maltose 1-phosphate (M1P) (Miah *et al.* 2013). M1P is expected to be acted upon by GlgE and GlgB to produce a glycogen-like α-glucan, given that the *treS/pep2*, *glgE* and *glgB* genes are clustered.

Other *Pseudomonas* spp. such as *P. fluorescens* and *P. syringae* possess an additional homologue of an unfused *treS* gene emphasising the importance of trehalose production in these organisms. Interestingly, this additional homologue is absent in *P. baetica*. *P. fluorescens* and *P. syringae* likely benefit from being able to control the synthesis of trehalose under more varied conditions.

We hypothesised that *P. baetica* was capable of causing disease in marine fish because it had greater salt tolerance than other pseudomonads; however, this proved to be incorrect. Our data show that all of the pseudomonads tested here are capable of growth in NaCl concentrations relevant to the marine environment (see Figures S8 and S9, Supporting Information).

In addition to salt tolerance experiments, *P. baetica* and *P. fluorescens* were also tested for their ability to grow in minimal media with the addition of Instant Ocean which has similar levels of most trace elements, nutrients and alkalinity to sea water samples (Fig. 3). *Pseudomonas baetica* grows well in Instant Ocean while *P. fluorescens* has minimal growth suggesting that *P. baetica* may have some adaptation to the marine environment. Analysis

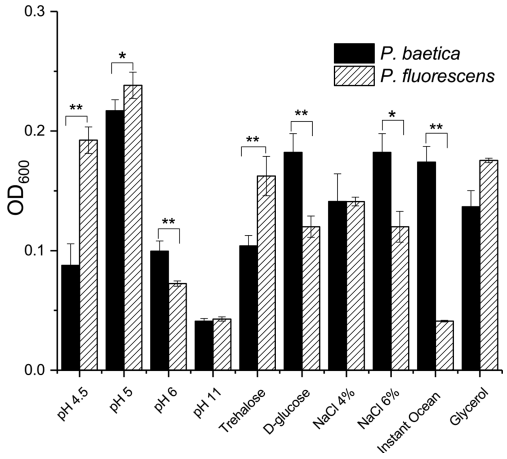


Figure 3. Stress tolerance and growth characteristics of *P. baetica* a390T. Graph of end point growth measured as OD_{600} after 24 h incubation at 30°C in *Pseudomonas* minimal media (L-glutamine 5%, K_2HPO_4 1.5%, $MgSO_4$ 0.2%, 20mM glycerol or other carbon source as indicated) with pH or salt concentration adjusted as indicated. Data represented mean ± SD. $^*P < 0.05$, $^{**}P < 0.01$ (determined by two-way *t*-test).

of both genomes using SEED viewer showed that *P. baetica* has a homologue of the *OsmY* gene which has a protein modification function during osmotic stress whereas *P. fluorescens* does not (Métris, George and Ropers 2017).

Iron uptake genes in *P. baetica*

The *P. baetica* genome was searched for genes coding for siderophore biosynthesis and uptake proteins as well as those involved in Fe^{2+} uptake. As expected, genes for the biosynthesis and uptake of the fluorescent siderophore pyoverdine are present (34 in total) and are scattered in four genomic loci as it is usually the case for fluorescent pseudomonads (Ravel and Cornelis 2003). *Pseudomonas baetica* also produces a second siderophore/metallophore, yersiniabactin, which was described in *Yersinia pestis* and in other Enterobacteriaceae (Chaturvedi *et al.* 2012). Yersiniabactin shares some biosynthesis steps with pyochelin, the second siderophore of *P. aeruginosa* and is able to bind copper with high affinity. It has been also described in some *P. syringae* and *P. avellanae* strains (Jones and Wildermuth 2011; Marcelletti and Scortichini 2015). Interestingly, while most pyoverdine genes have their closest homologs in *P. koreensis*, all, but one, yersiniabactin genes of *P. baetica* show a close identity with the yersiniabactin genes from *P. azotoformans*, which could be indicative of horizontal gene transfer. The genome contains 27 genes coding for a TonB-dependent receptor, including the *fpvA* gene for the ferripyoverdine uptake, and two *fyuA* genes for the uptake of ferri-yersiniabactin. However, two of these genes are truncated and a probable frameshift is evident for one receptor gene. Noteworthy is the presence of six genes annotated as coding for a TonB protein. As a matter of comparison, only three are present in *P. aeruginosa* (Cornelis, Matthijs and van Oeffelen 2009). Another surprising discovery is the absence of *feoABC* genes for the uptake of Fe^{2+} in *P. baetica*. This function is probably secured by the products of the *efeUOAB* operon, first described in *E. coli* O157:H7, which is involved in the uptake of ferrous iron in the periplasm where it gets re-oxidized (Cao *et al.* 2007).

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

DISCUSSION

Whole-genome sequencing has become ubiquitous in microbiology and access to the technology is now widespread. Here, we show that combining two relatively cheap and user-friendly methods (Ion Torrent Semiconductor and Oxford Nanopore MinION sequencing) enables genome assembly of a relatively large bacterial genome to a finishable state. Our phylogenetic analyses show that *P. baetica* is closely related to *P. koreensis* in the *P. fluorescens* group. However, our data show that the *P. fluorescens* group taxonomy is in need of review. Given the close relationship of *P. baetica* a390T to *P. irchel* 3e19, we propose that 3e19 is a strain of *P. baetica*. Although both are aquatic isolates, Irchel 3e19 was isolated from fresh water and *P. baetica* a390T is marine (López *et al.* 2012; Butaitė *et al.* 2017). Our data show that the ability to grow in marine salt concentrations is common in the genus *Pseudomonas*. This, combined with the origin of Irchel 3e19 suggests that *P. baetica* a390T is not a marine-restricted organism and is likely to be an opportunist pathogen of marine fish, possibly associated with high fish densities in aquaculture facilities. On the other hand, the presence of a T3SS IN BOTH *P. baetica* A390T and Irchel 3E19 (Prof. Rolf Kummerli, personal communication) suggests that these organisms are not merely benign environmental strains. The *P. baetica* genome shows evidence of genome plasticity and contains a variety of genomic islands and prophages as is the case with other pseudomonads including *P. aeruginosa* (Klockgether *et al.* 2011). The diverse iron acquisition and membrane transport systems encoded in the *P. baetica* genome are consistent with an organism that is highly adaptable to a range of environments. We find no genomic evidence of known QS systems found in other pseudomonads, but our observations that *P. baetica* is capable of protease secretion and swarming suggest that it may engage in social behaviour.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSLE* online.

FUNDING

Work in NPT's lab has been funded by Biotechnology and Biological Sciences Research Council grant BB/K019600/1 and Chief Scientists Office grant TCS/16/24. AB is supported by a PhD Studentship funded by the Industrial Biotechnology Innovation Centre and the University of Strathclyde. TS was supported by a PhD studentship funded by the Iraqi Government. MG was supported by a postdoctoral fellowship from FWO Vlaanderen (12M4618N). CL was supported by an SB PhD fellowship from FWO Vlaanderen (1S64718N). JB was supported by a grant from Fundação para a Ciência e a Tecnologia (SFRH/BD/104095/2014). NPT would like to personally thank all of the contributors to this paper who have made this novel collaboration an open and pleasurable experience. AM and ED were supported by the Biotechnology and Biological Sciences Research Council-funded South West Biosciences Doctoral Training Partnership (BB/M009122/1).

Conflict of interest. None declared.

REFERENCES

Andrews S. *FastQC A Quality Control tool for High Throughput Sequence Data*. 2014. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arahal DR. Whole-genome analyses: average nucleotide identity. *Method Microbiol* 2014;**41**:103–22.

Arndt D, Grant JR, Marcu A *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;**44**:W16–21.

Bankevich A, Nurk S, Antipov D *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.

Bertelli C, Laird MR, Williams KP *et al.* IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;**45**:W30–5.

Blin K, Wolf T, Chevrette MG *et al.* antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* 2017;**45**:W36–41.

Bravo JI, Lozano GL, Handelsman J. Draft Genome sequence of *Flavobacterium johnsoniae* CI04, an isolate from the soybean rhizosphere. *Genome Announc* 2017;**5**:e01535–16.

Butaitė E, Baumgartner M, Wyder S *et al.* Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater Pseudomonas communities. *Nat Commun* 2017;**8**:414.

Cao J, Woodhall MR, Alvarez J *et al.* EfeUOB (YcdNOB) is a tripartite, acid-induced and CpxAR-regulated, low-pH Fe 2+ transporter that is cryptic in Escherichia coli K-12 but functional in E.?coli O157:H7. *Mol Microbiol* 2007;**65**:857–75.

Castañeda-García A, Rodríguez-Rojas A, Guelfo JR *et al.* The glycerol-3-phosphate permease GlpT is the only fosfomycin transporter in *Pseudomonas aeruginosa*. *J Bacteriol* 2009;**191**:6968–74.

Catel-Ferreira M, Marti S, Guillon L *et al.* The outer membrane porin OmpW of Acinetobacter baumannii is involved in iron uptake and colistin binding. *FEBS Lett* 2016;**590**:224–31.

Cárcamo-Oyarce G, Lumjiaktase P, Kümmerli R *et al.* Quorum sensing triggers the stochastic escape of individual cells from *Pseudomonas putida* biofilms. *Nat Commun* 2015;**6**:5945.

Chandra G, Chater KF, Bornemann S. Unexpected and widespread connections between bacterial glycogen and trehalose metabolism. *Microbiology* 2011;**157**:1565–72.

Chaturvedi KS, Hung CS, Crowley JR *et al.* The siderophore yersiniabactin binds copper to protect pathogens during infection. *Nat Chem Biol* 2012;**8**:731–6.

Chevalier S, Bouffartigues E, Bodilis J *et al.* Structure, function and regulation of *Pseudomonas aeruginosa* porins. *FEMS Microbiol Rev* 2017;**41**:698–722.

Connor TR, Loman NJ, Thompson S *et al.* CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom* 2016;**2**:e000086.

Cornelis P, Matthijs S, van Oeffelen L. Iron uptake regulation in *Pseudomonas aeruginosa*. *Biometals* 2009;**22**:15–22.

De Smet KA, Weston A, Brown IN *et al.* Three pathways for trehalose biosynthesis in mycobacteria. *Microbiology* 2000;**146**:199–208.

Djonović S, Urbach JM, Drenkard E *et al.* Trehalose biosynthesis promotes *Pseudomonas aeruginosa* pathogenicity in plants. *PLoS Pathog* 2013;**9**:e1003217.

Dorosky RJ, Yu JM, Pierson LS, III *et al.* *Pseudomonas chlororaphis* produces two distinct R-tailocins that contribute to bacterial competition in biofilms and on roots. *Appl Environ Microb* 2017;**83**:AEM.00706–17.

Dubern J-F, Coppoolse ER, Stiekema WJ *et al.* Genetic and functional characterization of the gene cluster directing the biosynthesis of putisolvin I and II in *Pseudomonas putida* strain PCL1445. *Microbiology* 2008;**154**:2070–83.

Eren E, Vijayaraghavan J, Liu J *et al.* Substrate specificity within a family of outer membrane carboxylate channels. *PLoS Biol* 2012;**10**:e1001242.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

European Committee for Antimicrobial Susceptibility Testing (EUCAST) of the European Society of Clinical Microbiology and Infectious Dieases (ESCMID). EUCAST Definitive Document E.DEF 3.1, June 2000: Determination of minimum inhibitory concentrations (MICs) of antibacterial agents by agar dilution. *Clin Microbiol Infect* 2000;**6**:509–15.

Fann M-C, Busch A, Maloney PC. Functional characterization of cysteine residues in GlpT, the glycerol 3-phosphate transporter of Escherichia coli. *J Bacteriol* 2003;**185**:3863–70.

Freeman BC, Chen C, Beattie GA. Identification of the trehalose biosynthetic loci of *Pseudomonas syringae* and their contribution to fitness in the phyllosphere. *Environ Microbiol* 2010;**12**:1486–97.

Ghequire MGK, de Mot R. Ribosomally encoded antibacterial proteins and peptides from Pseudomonas. *FEMS Microbiol Rev* 2014;**38**:523–68.

Ghequire MGK, de Mot R. The Tailocin tale: peeling off phage tails. *Trends Microbiol* 2015;**23**:587–90.

Gomila M, Peña A, Mulet M *et al.* Phylogenomics and systematics in Pseudomonas. *Front Microbiol* 2015;**6**:214.

Grant JR, Arantes AS, Stothard P. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics* 2012;**13**:202.

Gupta V, Haider S, Sood U *et al.* Comparative genomic analysis of novel Acinetobacter symbionts: a combined systems biology and genomics approach. *Sci Rep* 2016;**6**:29043.

Gurevich A, Saveliev V, Vyahhi N *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5.

Hauser AR. The type III secretion system of *Pseudomonas aeruginosa*: infection by injection. *Nat Rev Microbiol* 2009;**7**:654–65.

Hittle LE, Powell DA, Jones JW *et al.* Site-specific activity of the acyltransferases HtrB1 and HtrB2 in *Pseudomonas aeruginosa* lipid A biosynthesis. *Pathog Dis* 2015;**73**:ftv053.

Huang Y, Lemieux MJ, Song J *et al.* Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli. *Science* 2003;**301**:616–20.

Jamet A, Nassif X. New players in the toxin field: polymorphic toxin systems in bacteria. *mBio* 2015;**6**:e00285–15.

Joh S-J, Ahn E-H, Lee H-J *et al.* Bacterial pathogens and flora isolated from farm-cultured eels (Anguilla japonica) and their environmental waters in Korean eel farms. *Vet Microbiol* 2013;**163**:190–5.

Jones AM, Wildermuth MC. The phytopathogen *Pseudomonas syringae* pv. tomato DC3000 has three high-affinity iron-scavenging systems functional under iron limitation conditions but dispensable for pathogenesis. *J Bacteriol* 2011;**193**:2767–75.

Kahan FM, Kahan JS, Cassidy PJ *et al.* The mechanism of action of fosfomycin (phosphonomycin). *Ann NY Acad Sci* 1974;**235**:364–86.

Kerry J, Hiney M, Coyne R *et al.* Fish feed as a source of oxytetracycline-resistant bacteria in the sediments under fish farms. *Aquaculture* 1995;**131**:101–13.

Keshavarz-Tohid V, Taheri P, Muller D *et al.* Phylogenetic diversity and antagonistic traits of root and rhizosphere pseudomonads of bean from Iran for controlling *Rhizoctonia solani*. *Res Microbiol* 2017;**168**:760–72.

Klockgether J, Cramer N, Wiehlmann L *et al.* *Pseudomonas aeruginosa* genomic structure and diversity. *Front Microbiol* 2011;**2**:150.

Knirel YA, Bystrova OV, Kocharova NA *et al.* Review: conserved and variable structural features in the lipopolysaccharide of *Pseudomonas aeruginosa*. *J Endotoxin Res* 2006;**12**:324–36.

Kucharska I, Seelheim P, Edrington T *et al.* OprG harnesses the dynamics of its extracellular loops to transport small amino acids across the outer membrane of *Pseudomonas aeruginosa*. *Structure* 2015;**23**:2234–45.

Lam JS, Taylor VL, Islam ST *et al.* Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide. *Front Microbiol* 2011;**2**:118.

Lamarche MG, Déziel E. MexEF-OprN efflux pump exports the *Pseudomonas quinolone* signal (PQS) precursor HHQ (4-hydroxy-2-heptylquinoline). *PLoS One* 2011;**6**:e24310.

Law CJ, Almqvist J, Bernstein A *et al.* Salt-bridge dynamics control substrate-induced conformational change in the membrane transporter GlpT. *J Mol Biol* 2008;**378**:828–39.

Leal JF, Henriques IS, Correia A *et al.* Antibacterial activity of oxytetracycline photoproducts in marine aquaculture's water. *Environ Pollut* 2017;**220**:644–9.

Li XZ, Zhang L, Poole K. Role of the multidrug efflux systems of *Pseudomonas aeruginosa* in organic solvent tolerance. *J Bacteriol* 1998;**180**:2987–91.

Llanes C, Köhler T, Patry I *et al.* Role of the MexEF-OprN efflux system in low-level resistance of *Pseudomonas aeruginosa* to ciprofloxacin. *Antimicrob Agents Ch* 2011;**55**:5676–84.

López JR, Diéguez AL, Doce A *et al.* *Pseudomonas baetica* sp. nov., a fish pathogen isolated from wedge sole, *Dicologlossa cuneata* (Moreau). *Int J Syst Evol Microbiol* 2012;**62**:874–82.

López JR, Lorenzo L, Marcelino-Pozuelo C *et al.* *Pseudomonas baetica*: pathogenicity for marine fish and development of protocols for rapid diagnosis. *FEMS Microbiol Lett* 2017;**364**:fnw286.

Marcelletti S, Scortichini M. Comparative genomic analyses of multiple *Pseudomonas* strains infecting *Corylus avellana* trees reveal the occurrence of two genetic clusters with both common and distinctive virulence and fitness traits. *PLoS One* 2015;**10**:e0131112.

Métris A, George SM, Ropers D. Piecewise linear approximations to model the dynamics of adaptation to osmotic stress by food-borne pathogens. *Int J Food Microbiol* 2017;**240**:63–74.

Miah F, Koliwer-Brandl H, Rejzek M *et al.* Flux through trehalose synthase flows from trehalose to the alpha anomer of maltose in mycobacteria. *Chem Biol* 2013;**20**:487–93.

Michel-Briand Y, Baysse C. The pyocins of *Pseudomonas aeruginosa*. *Biochimie* 2002;**84**:499–510.

Miranda CD, Kehrenberg C, Ulep C *et al.* Diversity of tetracycline resistance genes in bacteria from Chilean salmon farms. *Antimicrob Agents Ch* 2003;**47**:883–8.

Ondov BD, Treangen TJ, Melsted P *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;**17**:132.

Pascual J, García-López M, Bills GF *et al.* *Pseudomonas granadensis* sp. nov., a new bacterial species isolated from the Tejeda, Almijara and Alhama Natural Park, Granada, Spain. *Int J Syst Evol Microbiol* 2015;**65**:625–32.

Perkel JM. How scientists use Slack. *Nature* 2016;**541**:123–4.

Poole K. *Pseudomonas aeruginosa*: resistance to the max. *Front Microbiol* 2011;**2**:65.

Rambaut A. *FigTree v1. 4.* 2012. http://tree.bio.ed.ac.uk/software/figtree/ (7 December 2017, date last accessed).

Ramírez-Bahena M-H, Cuesta MJ, Flores-Félix JD *et al.* *Pseudomonas helmanticensis* sp. nov., isolated from forest soil. *Int J Syst Evol Microbiol* 2014;**64**:2338–45.

Ravel J, Cornelis P. Genomics of pyoverdine-mediated iron uptake in pseudomonads. *Trends Microbiol* 2003;**11**:195–200.

Richter M, Rosselló-Móra R, Oliver Glöckner F *et al.* JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 2016;**32**:929–31.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021

Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9.

Sharp C, Bray J, Housden NG *et al.* Diversity and distribution of nuclease bacteriocins in bacterial genomes revealed using Hidden Markov Models. *PLoS Comput Biol* 2017;**13**:e1005652.

Smith EW, Zhang X, Behzadi C *et al.* Structures of *Pseudomonas aeruginosa* LpxA reveal the basis for its substrate selectivity. *ACS Publications* 2015;**54**:5937–48

Vaser R, Sović I, Nagarajan N *et al.* Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737–46.

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

Wick RR, Judd LM, Gorrie CL *et al.* Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;**13**:e1005595.

Wick RR, Schultz MB, Zobel J *et al.* Bandage: interactive visualization of de novo genome assemblies: Fig. 1. *Bioinformatics* 2015;**31**:3350–2.

Xu S, Fu Z, Zhou Y *et al.* Mutations of the Transporter proteins GlpT and UhpT confer fosfomycin resistance in *Staphylococcus aureus*. *Front Microbiol* 2017;**8**:914.

Zhang D, de Souza RF, Anantharaman V *et al.* Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct* 2012; **7**:18.

Downloaded from https://academic.oup.com/femsle/article/365/9/fny069/4951603 by John Innes Centre user on 26 June 2021