UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

# Comparative genomic analyses of cyanobacteria

Carolina de Castro Barbosa Rodrigues Barata

**Mestrado em Biologia Evolutiva e do Desenvolvimento**

Dissertação orientada por:
Professor Octávio Paulo e Doutora Elsa Dias

2017

# Acknowledgements

I would like to thank my supervisor Professor Octávio Paulo for giving me the opportunity and all the means to work on such an interesting topic. This project has taught me a lot and I enjoyed every minute of it. Data acquisition is hard work and pursuing a career as an Evolutionary Biologist has proved me exactly that.

I thank my supervisor Doctor Elsa Dias, whose kindness and patience showed me that laboratory work is nothing to be afraid of. Coping with disappointing results is far easier if you work alongside good people.

I would also like to thank everyone at Cobig2, especially Francisco Pina-Martins. Everybody's resilience was such an inspiration for this past year.

Finally, I thank my family and friends, who have always cheered me up in times of absolute despair and chaos.

# Abstract

Cyanobacteria are photosynthetic organisms that can be found in water bodies worldwide. Cyanobacterial lineage evolution resulted in numerous colour, shape and colony-forming phenotypes. Most cyanobacteria produce several toxins, but only a few are capable of nitrogen ($N_2$) fixation. Most interestingly, these bacteria exhibit antimicrobial activity, as well as antibiotic resistance phenotypes. Therefore, it has been suggested that cyanobacteria might harbour several antibiotic resistance (AR) genes. Consequently, it has been hypothesized that these bacteria might play a major role in the dissemination of antibiotic resistance phenotypes in microbial communities. Whole-genome sequencing is necessary to fully characterise the genetic basis of such AR phenotypes. In particular, long-read third-generation sequencing methods might help resolve the genomic structure of AR gene containing regions. Horizontally transferred AR genes are often flanked by highly repetitive sequences, not completely resolved by high-throughput next-generation sequencing technologies. Here, we report the use of the portable MinION (Oxford Nanopore Technologies, UK) sequencing device to obtain the genome sequences of six strains of cyanobacteria. Said biological isolates were collected at different time points in freshwater bodies across Portugal. We obtained genome scaffolds for two *Microcystis aeruginosa* strains. The LMECYA7 scaffold resulted from a hybrid assembly using not only MinION data but also Illumina paired-end reads. The LMECYA167 genome scaffold represented more than 68% of the reference genome and was entirely built from MinION 1D reads. For both strains, more than one hundred homologous AR gene sequences were found. Among these, fluoroquinolone, beta-lactam and sulfonamide resistance-associated genes were present. Moreover, genome annotation might have unveiled the genetic basis of trimethoprim resistance in the aforementioned cyanobacterial strains. The presence of an alternative folate pathway enzyme (thymidylate synthase, thyX) could fully account for such trimethoprim resistance phenotype. In summary, MinION sequencing allowed us to not only find several homologous AR genes, but also pinpoint the genetic mechanism that is responsible for trimethoprim resistance in two strains of cyanobacteria. Lastly, the process of gDNA extraction requires further optimisation, in order to obtain maximum MinION sequencing yield.

**Keywords:** cyanobacteria, genomics, MinION

# Resumo

O presente trabalho teve como objectivo estudar o genoma de 6 (seis) estirpes de cianobactérias recolhidas de locais distintos de Portugal continental. Durante a duração do projecto, proposemo-nos a extrair e quantificar DNA genómico das estirpes mencionadas para, então, procedermos à sua sequenciação. A posterior análise dos dados obtidos focou-se em procurar genes conhecidos de resistência a antibióticos bem como outros elementos genómicos que possam conferir fenótipos de resistência a substâncias antimicrobianas.

As cianobactérias são um grupo de procariotas muito estudado quanto à sua prevalência em reservatórios de água para consumo humano e animal. A sua importância prende-se no facto de diversas linhagens de cianobactérias produzirem compostos tóxicos, nomeadamente microcistinas. Estes compostos são nocivos para células eucarióticas, podendo mesmo conduzir à morte dos indivíduos que as ingerem. Para além disto, este grupo de organismos apresenta capacidade de fotossíntese oxigénica, contribuindo para a produção de oxigénio a partir de compostos orgânicos.

Mais recentemente, o papel das cianobactérias como parte integrante das comunidades de microorganismos que habitam ambientes aquáticos foi repensado. Estas bactérias proliferam tanto em água doce como em água salgada e contribuem para o *pool* genético das comunidades em que se encontram. As cianobactérias são capazes de adquirir e partilhar tanto genes como outros elementos genómicos, nomeadamente através de transferência horizontal. A fluidez dos genomas bacterianos sugere que, tal como os restantes genes, os genes de resistência a antibióticos são partilhados pela comunidade microbiana. Este *pool* de genes de resistência denomina-se resistoma. Na medida em que as cianobactérias apresentam fenótipos de resistência a antibióticos, o seu papel no resistoma destas comunidades começa a ser alvo de estudo.

De forma a caracterizar o resistoma de uma comunidade microbiana, é necessário conhecer todos os genes que possam conferir resistência a determinados antibióticos. Para tal, é essencial sequenciar o genoma completo dos indivíduos recolhidos. A Estela Sousa e Silva Algae Culture Collection (ESSACC), actualmente localizada no Instituto de Saúde Doutor Ricardo Jorge, e a Blue Biotechnology and Ecotoxicology Culture Collection (LEGE), localizada no Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR), permitem-nos iniciar um estudo de caracterização do resistoma de microorganismos de água doce em Portugal. Estas colecções incluem amostras de reservatórios naturais, estações de tratamento de águas residuais e barragens. As amostras caracterizam-se por pertencer a diversas espécies com fenótipos distintos, nomeadamente no que diz respeito a coloração, forma e produção de colónias.

Após um estudo prévio de caracterização dos fenótipos de resistência a antibióticos de numerosas amostras, 6 estirpes foram escolhidas para este trabalho. Esta selecção teve em conta fenótipos de resistência contrastantes, ou seja, a escolha teve por base estirpes que apresentassem susceptibilidades diferentes para dados grupos de antibióticos. Ainda assim, todas as estirpes apresentaram resistência ao trimetoprim e ao ácido nalidíxico. Duas estirpes pertencem à espécie *Microcystis aeruginosa* (LMECYA7 e LMECYA167), outras duas são classificadas como *Plantothrix agardhii* (LMECYA269 e LMECYA280) e as restantes duas como *Planktothrix mougeotii* (LEGE06226 e LEGE06233).

Quanto à sequenciação, optámos por utilizar o método de terceira geração MinION (Oxford Nanopore Technologies, ONT, UK). O pequeno aparelho de sequenciação funciona através de uma ligação USB 3.0 a qualquer computador portátil ou *desktop*. A sequenciação dá-se através de um poro que está ligado a um transdutor de sinal eléctrico. A molécula de DNA passa através do poro e a corrente medida é transformada numa sequência de bases. Este método tem inúmeras vantagens, nomeadamente:

- Permite obter *reads* longas, na ordem dos milhares de pares de bases (*base pairs*, bp), o que não é possível através dos métodos convencionais de segunda geração;
- Permite obter sequências a partir de DNA não amplificado, ou seja, reduz-se a necessidade aumentar a quantidade de DNA através de PCR;
- Permite visualizar o processo de sequenciação em tempo real, no que diz respeito ao rendimento de cada corrida.

Para além das vantagens supra referidas, uma das modalidades deste método permite preparar a biblioteca genómica para sequenciação em apenas alguns minutos. Contudo, o processo de obtenção do DNA constitui o passo limitante no que diz respeito a uma sequenciação bem sucedida. A qualidade e quantidade de DNA utilizado para a preparação da biblioteca têm de obeceder a critérios rigorosos. O tamanho dos fragmentos que constituem a biblioteca é um outro factor limitante, no sentido em que, quanto mais fragmentado estiver o DNA, mais pequenas serão as *reads* obtidas. Portanto, menor será a resolução do genoma que os dados vão permitir.

Em suma, realizámos duas rondas de sequenciação com MinION. Na primeira, procedemos a uma corrida de teste, na qual sequenciámos DNA de fago lambda, fornecido pela ONT. Sequenciámos ainda duas bibliotecas de DNA de cianobactérias, pertencentes às estirpes LMECYA167 e LEGE06233. O DNA genómico de ambas as estirpes foi obtido através da utilização de um *kit* de extracção. Relativamente à segunda ronda de sequenciação, extraímos o DNA das 6 estirpes já mencionadas através da realização de um protocolo de fenol-clorofórmio. Este protocolo tem como objectivo a obtenção de material genético de elevado peso molecular.

No final de ambas as rondas de sequenciação, conseguimos obter *scaffolds* dos genomas das duas estirpes de *M. aeruginosa* em estudo, isto é, das estirpes LMECYA7 e LMECYA167. Os dados obtidos para as restantes estirpes ficaram aquém das expectativas, no sentido em que não permitiram obter cobertura significativa de nenhum dos genomas. A análise dos dados de todas as sequenciações revelou que a qualidade das *reads* obtidas foi baixa. Os dados foram, então, filtrados para que apenas as *reads* de melhor qualidade fossem assembladas e alinhadas aos respectivos genomas de referência. Porém, para todas as estirpes do género *Planktothrix*, a quantidade de *reads* mapeada contra a referência e, consequentemente, a cobertura obtida foram irrisórias.

Quanto aos *scaffolds* dos genomas das estirpes de *M. aeruginosa*, ambas produziram uma sequência *consensus* maior que metade do genoma de referência. Assim sendo, procedemos à análise dos potenciais genes de resistência a antibióticos presentes nas sequências referidas. Tanto em LMECYA7 como em LMECYA167, foram encontrados potenciais homólogos de genes associados a resistências. Na maioria dos casos, estes genes pertencem a componentes proteicos de bombas de efluxo. Alguns dos outros possíveis genes homólogos encontrados surgem associados a resistência a beta-lactâmicos, quinolonas

e sulfonamidas. Genes de resistência a tetraciclina também estão presentes na lista de resultados. Finalmente, a anotação dos *scaffolds* obtidos parece ter permitido entender a base genética da resistência ao trimetoprim. Foram encontrados, nas sequências de ambas as estirpes, genes para um enzima alternativo da via metabólica dos folatos. Esta via é essencial à sobrevivência destes microorganismos e o trimetoprim actua de forma a inibir um dos enzimas da via. Com a presença de um enzima alternativo, a resistência ao antibiótico é assegurada.

Em conclusão, os dados obtidos através de sequenciação de terceira geração permitiram obter *scaffolds* de dois genomas de cianobactérias. O processo de sequenciação foi simples, mas limitado pelo relativo sucesso do processo de extracção. A análise dos dados também foi limitada pela actualização constante dos softwares em actual desenvolvimento. Em última análise, foi possível detectar sinais de adaptação no que diz respeito à evolução de fenótipos de resistência a antibióticos nas estirpes em estudo.

**Palavras-chave:** cianobactérias, genómica, MinION

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Evolutionary genetics and adaptation

Understanding causes and patterns of evolution across the tree of life can prove difficult. However, researchers are now able to accurately describe the evolutionary history of numerous groups of organisms. Tracking events that happened in the past of a given lineage is possible by analysing its genome. The evolution of genomes is shaped by mutation, selection and drift. Third generation sequencing methods are powerful tools for the study of such phenomena.

It has been long since researchers found that selection can contribute to the evolution of populations only if there is variance in fitness among the individuals (Fisher, 1930). Different combinations of advantageous and deleterious alleles generate such variance (Felsenstein, 1974). Accordingly, selection can cause allele frequency shifts in order to allow populations to adapt. However, in finite populations, genetic draft has to overcome the effects of genetic drift to be successful at causing adaptation. Evidence of population adaptation has been found in both model and non-model organisms from single gene analyses (Mundy, 2005; Hoekstra et al., 2006) to whole-genome scans (Bennett and Lenski, 2007; Sattath et al., 2011; Silva et al., 2015).

## 1.2 Cyanobacteria as model organisms

This project aims at detecting signatures of ecological adaptation by studying the main genomic features of six strains of cyanobacteria collected in different waterbodies across Portugal. Cyanobacteria are a group of photosynthetic bacteria. Their origin is thought to have occurred approximately 2.5 billion years ago (Altermann and Kazmierczak, 2003). Ever since, cyanobacteria have evolved to occupy numerous ecological niches exhibiting a plethora of morphological and physiological adaptations, as well as specific genomic features (Hess, 2011). These microorganisms range from unicellular to colonial and filamentous forms. They are even capable of nitrogen ($N_2$) fixation in specialised cells called heterocysts (Karl et al., 2002). Only filamentous colonies produce such cells. The ability to fix atmospheric $N_2$ is referred to as diazotrophy, which was be proposed to be an ancestral trait of the prokaryote lineage (Stucken et al., 2010). In a 2006 study, Tomitani et al. found that diazotrophy is polyphyletic in 20 spp of cyanobacteria, indicating that this trait was present in the last common ancestor of such lineages. Moreover, cyanobacteria are known to thrive in both freshwater and marine ecosystems and to occupy benthic and planktonic niches.

Cyanobacteria are found in water bodies worldwide associated with harmful algal blooms (HABs). HABs consist of not only cyanobacteria but also microscopic algae. These particular blooms of microorganisms have been extensively studied since they cause both human and animal disease and even death. Such toxic blooms have been described for more than 100 years (Francis, 1878). Although cyanobacterial blooms can be a human and livestock health hazard, particular secondary metabolites produced by cyanobacteria bear great pharmaceutical and biotechnological potential (Hess, 2011). Accordingly, and as primary producers, cyanobacteria play a major role in aquatic ecosystems. Understanding the evolutionary and ecological history of the cyanobacterial lineage using novel genome sequencing methods is, thus, increasingly compelling.

Cyanobacteria are capable of producing different types of toxins, such as neurotoxins and hepatotoxins. Of the latter, microcystins (MCs) are a well-studied group of toxic cyclic peptides (Tillett et al., 2000) which are produced by a diverse range of cyanobacteria (Rinehart et al., 1994; Sivonen, 1996). MCs are potent inhibitors of eukaryotic protein phosphatases. These toxins are actively transported into vertebrate hepatocytes causing permanent damage (Eriksson et al., 1990). MCs and other cyanobacterial peptides are produced via nonribosomal peptide synthases (NRPSs). NRPS gene clusters have been studied and characterised in several genera of cyanobacteria (Calteau et al., 2014), namely in *Planktothrix* (Rounge et al., 2009).

The microcystin-producing *mcy* gene cluster has been described in at least three orders of cyanobacteria, Chroococales, Nostocales and Oscillatoriales (Jungblut and Neilan, 2006). The *mcy* gene cluster of a *Microcystis aeruginosa* strain (Chroococales) was characterised in detail in 2000 by Tillett et al. Later, Kurmayer et al. (2005) studied said cluster in 17 *Planktothrix* strains (Oscillatoriales). Moreover, researchers have found evidence of purifying selection on the *mcy* genes across the cyanobacterial lineage, namely low $K_A/K_S$ (Rantala et al., 2004) and $d_N/d_S$ ratios (Kurmayer and Gumpenberger, 2006). In a more recent study, Kurmayer et al. (2015) performed a phylogenetic analysis of 138 *Planktothrix* strains from distinct geographical origins. All the strains contained at least remnants of the cluster and a non-polarised McDonald-Kreitman test revealed no evidence of selection acting on *mcy* genes.

Based on evidence for fitness effects of MC production, authors have suggested several putative roles for these toxins. In both *Planktothrix* (Briand et al., 2008) and *Microcystis* (Schatz et al., 2007) genera, the benefits of producing the toxin have been shown to outweigh the costs under growth-limiting conditions. Moreover, under limiting light and temperature conditions, MC-producing strains had greater fitness. Thus, a strain that produces MCs has a selective advantage that allows it to thrive in its microbial community. Accordingly, MCs were suggested to play a role in light adaptation processes (Hesse et al., 2001) and siderophoric scavenging of trace metals (Rantala et al., 2004), as well as intraspecies communication (Schatz et al., 2007) and quorum-sensing (Kaebernick and Neilan, 2001). MC production might also play a part as a feeding deterrent or in cell-to-cell signaling and gene regulation. Contrastingly, in *P. agardhii* strains, researchers were able to associate insertion sequence (IS) elements, which are a type of transposable element (TE), to the inactivation or recombination of cyanotoxin production genes (Christiansen et al., 2014). Although the link between toxicity and specific environmental cues remains unclear, the role of natural selection seems to be key in the diversification of toxic and non-toxic strains of cyanobacteria.

The genomes of cyanobacteria are highly variable and plastic (Rocha, 2006). This is mostly due to a phenomenon known as Horizontal Gene Transfer (HGT). HGT is mostly caused by three distinct processes: transformation, when bacteria uptake environmental DNA; conjugation of DNA via cell-cell communication; and transduction, as a result of infection by lysogenic phages. In an effort to find a cyanobacterial core genome, Shi and Falkowski (2008) found that more than 52% of orthologous genes were susceptible to HGT within 13 species of cyanobacteria. In a 2003 study, Rocap et al. were also able to find evidence of HGT in 2 strains of *Prochlorococcus*. Similarly, 11% of genes were found to be the result of recent transfer events among 12 strains of *M. aeruginosa* (Humbert et al., 2013). HGT events can range from gene fragments, to complete genes and even whole operons. For instance, Tooming-Klunderud et al. (2013) found evidence of HGT of the phycoerythrin gene cluster between

*Planktothrix* strains through homologous recombination. Thus, HGT allows cyanobacterial populations to adapt quicker to varying environmental conditions.

Recently, environmental microbial communities have been suggested to be important reservoirs for antibiotic resistance (AR) genes (Manageiro et al., 2014; Amos et al., 2014). It has also been suggested that cyanobacteria might play a role in the environmental resistome, i.e. the pool of AR genes found in such microbial communities. Accordingly, some species of cyanobacteria exhibit antibacterial activity (Wright, 2007) as well as resistance to a few antibiotics, such as ampicilin and penicilin (Prasanna et al., 2010). Furthermore, in 2015, Dias et al. studied antibiotic susceptibility of cyanobacteria belonging to the Estela Sousa e Silva Algae Culture Collection (ESSACC). Susceptibility was assessed for *Aphanizomenon gracile*, *Chrisosporum bergii* and *Planktothrix agardhii* as well as 9 *Microcystis aeruginosa* strains. Cell growth was followed over time. All strains exhibited highest susceptibility to $\beta$-lactams. Additionally, no isolates were susceptible to trimethoprim or nalidixic acid.

Consequently, researchers are working towards finding the genetic basis of AR in cyanobacteria. Plasmids and other mobile genetic elements, such as TEs, have been proposed to determine AR (Bennett, 2009; Chen et al., 2008). Said genomic elements are required for HGT to happen. Thus, cyanobacteria must bear AR genes because both plasmids and TEs have been successfully found in cyanobacterial genomes (Christiansen et al., 2008). Furthermore, HGT has been shown to allow for the quick spread of numerous genomic features between strains of bacteria (Davies, 1994; Marti et al., 2013; Huddleston, 2014; Jiao et al., 2017). However, scientists are yet to accurately describe how does this process facilitate the spread of AR genes among cyanobacteria.

## 1.3 Third generation MinION sequencing technology

In the early 1990's, bacterial whole-genome sequencing revealed thousands of new genes as well as a major role for HGT in generating genetic diversity. Comparative genomics began to tell core and pan genomes apart by reporting genes present in all strains within a taxonomic group versus the entire set of genes that characterise such taxon. Later on, high-throughput next-generation sequencing methods generated SNP-based phylogenies of bacterial lineages. Moreover, reconstructing transmission chains and characterising within-patient pathogen diversity became feasible. As of 2010, third-generation real-time sequencing technologies, such as Pacific Biosciences' SMRT and Oxford Nanopore Technologies' MinION, promised to revolutionise whole-genome data acquisition (Lu et al., 2016). Both sequencing devices are able to produce much longer reads in comparison to next-generation technologies. Third-generation sequencing reads are tens of thousand base-pairs long. Moreover, no fragment amplification is required and, consequently, no sequencing biases from PCR amplification are produced. Additionally, sequencing time is drastically reduced from days to hours.

MinION is a highly portable device that connects to a laptop or desktop computer via a USB 3.0 port. DNA is loaded into a consumable flow cell and directly sequenced. Each flow cell contains up to 2048 nanopore channels connected to 512 signal amplifiers. All nanopores are inserted in a silicon membrane and contain a platinum electrode at the base of the well. The nanopore signal is processed by device-management software MinKNOW. During an experiment, a customisable Python script controls the device and the flow cell. A DNA molecule is sequenced as it goes through the pore. In turn, the amplifier channel records the current signal over time. Varying voltage should be the result of different nucleotides passing through the nanopore. In contrast to all other sequencing technologies, MinION is able to produce reads of virtually unlimited length. Since a DNA molecule is directly sequenced as it goes through the channel, read length is sole function of the molecule's initial size. Thus, ONT's nanopore technology might help resolve the sequence of highly repetitive genomic regions.

Despite numerous advantages, MinION whole-genome sequencing is hindered by high error rates and homopolymer under-representation. Because nanopore readings exhibit high stochastic variability, MinION error rates are not negligible. Rate estimation is usually based on sequence identity to either a reference genome or Illumina reads. Over the years, flowcell chemistry improvements and management software updates have resulted in greater sequence identity. Reported raw read error rates are as high as 38.2%, for R6 chemistry flowcells (Laver et al., 2015), and as low as 8%, for R9.4 chemistry flowcells (Benítez-Páez and Sanz, 2017). Other studies estimate error rates between 25-17.2% and 14.8-11% for R7.3 and R9 flowcells, respectively (Istace et al., 2016; Jansen et al., 2017). Moreover, MinION raw reads exhibit substantial misrepresentation of homopolymers, i.e. repeats of identical nucleotides (Lu et al., 2016). Basecalling software implements a maximum k-mer length. Therefore, it might overlook homopolymer sequences that are longer than such k-mer size. This could have an impact on genome size estimation, overall GC content and repetitive sequence resolution.

MinION data publication list is increasing exponentially since 2016. For instance, MinION sequencing has proved useful for metagenomic studies. In a 2016 paper, Edwards et al. used nanopore technology to characterise the microbiota of a High Arctic glacier. Additionally, Brown et al. (2017) tested the ability of correctly assigning taxonomy from MinION data in both synthetic bacterial communities and single species runs. Furthermore, several bacterial genomes, such as *Escherichia coli*'s (Jain et al., 2017) and *Staphylococcus aureus*' (Bayliss et al., 2017), were sequenced using MinION. Some eukaryotic genomes were also MinION sequenced, namely *Saccharomyces cerevisiae* yeast (Istace et al., 2016; Giordano et al., 2017). Notably, the genome sequences of a few animal and plant species have been obtained using nanopore technology, for example the European eel (*Anguilla anguilla*) (Jansen et al., 2017), wild tomato (*Solanum pennelli*) (Schmidt et al., 2017b) and thale cress (*Arabidopsis thaliana*) (Michael et al., 2017). Lastly, researchers have used MinION data to identify and characterise antibiotic resistance genes (Ashton et al., 2014; Ludden et al., 2017; Schmidt et al., 2017a).

This work aims to unveil signatures of population adaptation from genome sequencing data of six cyanobacterial strains, belonging to three species - *Microcystis aeruginosa*, *Planktothrix agardhii* and *Planktothrix mougeotii*. Using ONT's MinION third-generation sequencing technology, we focused on obtaining good quality whole-genome sequences. We worked towards identifying the main genomic features that allow different cyanobacterial strains to exhibit antibiotic resistance phenotypes.

# 2 Methodology

## 2.1 Cyanobacteria

The strains of cyanobacteria used throughout this work belong to the Estela Sousa e Silva Algae Culture Collection (ESSACC) from Instituto de Saúde Doutor Ricardo Jorge (Paulino et al., 2009), and the Blue Biotechnology and Ecotoxicology Culture Collection (LEGE) from Centro Interdisciplinar de Investigação Marinha e Ambiental (Martins et al., 2010). The strains were isolated from freshwater bodies across Portugal with numerous geographical/usage characteristics. All strains were phenotipically characterised and their phylogenetic relationship was previously inferred (Valério et al., 2009). Each isolate is maintained in Z8 medium as a monoalgal, free of eukaryotes, non-axenic stock culture. The laboratory culture chamber is kept in a 14/10 h L/D cycle (light intensity 16 ± 4 $\mu$Em$^{-2}$s$^{-1}$, approx.) at 20±1°C.

Four ESSACC and two LEGE strain genomes belonging to two different genera were sequenced during the course of this work. ESSACC strains are referred to as LMECYA strains. The LMECYA7 and LMECYA167 strains belong to the *Microcystis aeruginosa* species and were first isolated in the rural freshwater reservoirs of Montargil (1996) and Corgas (2001), respectively. The remaining four strains belong to the *Planktothrix* genus. LMECYA269 was collected from the Magos reservoir in 2009, while LMECYA280 was collected from the São Domingos reservoir in 2012. Both strains are *Planktothrix agardhii* isolates. Finally, LEGE06226 and LEGE06233 belong to the *Planktothrix mougeotii* species and were isolated from the Febros waste water treatment plant in 2006 and 2007, respectively.

The strains were chosen based according to antibiotic susceptibility phenotypes, particularly the Minimum Inhibitory Concentration (MIC). In this case, cell growth was determined by optical density measurements (450 nm) of the antibiotic exposed cultures, relatively to control cultures (not exposed to antibiotics). Furthermore, culture integrity was followed by microscopic examination and each MIC estimate corresponds to the absence of undamaged cyanobacterial cells. MICs were previously obtained for a set of at least 9 different antibiotics for all six strains (Dias et al., 2015). A detailed description of all antibiotics and corresponding MICs can be found in table 1. Lowest MICs correspond to highest susceptibility.

Table 1: MICs of all six strains of cyanobacteria. AMX: amoxicillin, CAZ: ceftazidime, CRO: ceftriaxone, KAN: kanamycin, GEN: gentamicin, STR: streptomycin, NOR: norfloxacin, NAL: nalidixic acid, TET: tetracycline, TMP: trimethoprim

| | MIC (mg/L) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Strain** | **AMX** | **CAZ** | **CRO** | **KAN** | **GEN** | **STR** | **NOR** | **NAL** | **TET** | **TMP** |
| LMECYA7 | 0.8 | 0.4 | 0.2 | 0.2 | 0.2 | NA | 0.05 | >1.6 | >0.8 | >1.6 |
| LMECYA167 | 0.2 | 0.8 | 0.8 | 0.2 | 0.2 | NA | 0.2 | >1.6 | >1.6 | >1.6 |
| LMECYA269 | 0.025 | 0.4 | 0.4 | 0.2 | 0.1 | 0.025 | 1.6 | >1.6 | 0.2 | >1.6 |
| LMECYA280 | 0.8 | 1.6 | 1.6 | 0.2 | 0.1 | 0.05 | >1.6 | >1.6 | 0.8 | >1.6 |
| LEGE06226 | >1.6 | 0.4 | 0.2 | 0.2 | 0.4 | 0.05 | >1.6 | >1.6 | 1.6 | >1.6 |
| LEGE06233 | >1.6 | 0.8 | 0.4 | 0.8 | 0.8 | 0.8 | >1.6 | >1.6 | >1.6 | >1.6 |

## 2.2 DNA extraction and sequencing

We used the Oxford Nanopore Techologies (ONT, UK) sequencing platform, MinION, to sequence the genomes of the six cyanobacterial strains previously chosen. Two sequencing rounds were performed. For the first, genomic DNA of strains LMECYA167 and LEGE06233 was extracted using the MO BIO Laboratories, Inc. PowerWater DNA Isolation Kit. The biomass for DNA extraction was obtained from 20 mL of cyanobacterial cultures in exponential growth phase were centrifuged for 5 minutes at 2000x*g*. Extracted DNA was concentrated using SpeedVac for an average of 35 minutes at medium heat. DNA quality and concentration were assessed using a Nanodrop spectrophotometer (ND-1000) and a Qubit 2.0 fluorometer (dsDNA High Sensitivity assay, Life Technologies). A maximum of 200 ng cyanobacterial DNA was used for library preparation. Sequencing library preparation was carried out using the Rapid Sequencing Kit SQK-RAD002 for the two LMECYA strains and the lambda phage DNA provided in the kit. The libraries were loaded into a MinION R9.5 flow cell on a MinION MK 1B controlled by MinKNOW software version 1.6.11. Lambda phage sequencing ran on a with a 6hr run plus Basecaller.py workflow. Both LMECYA167 and LEGE06233 gDNA was sequenced by running a 48-hour run plus Basecaller.py workflow. However, data acquisition was stopped after 12 and 18 hours for LMECYA167 and LEGE06233, respectively. Live base calling was performed using the ONT EPI2ME Agent version 2.47.537208. The flow cell was washed between sequencing runs using the ONT Wash Kit WSH-002 and kept at 4°C overnight.

For the second sequencing round, DNA of all six cyanobacterial strains was extracted using a high molecular weight DNA extraction protocol (Sambrook and Russell, 2001) modified by Jain et al. (2017). Cyanobacterial cells were collected from exponential growth phase cultures into Falcon tubes. Between 15-60 mL of cultured cells were spun for at least 10 minutes at 2000x*g* in order to pellet. The supernatant was removed and the cells were resuspended in 100 $\mu$L PBS. Ten (10) mL of TLB (10 mM Tris-HCl pH 8.0, 25 mM EDTA pH 8.0, 0.5% (w/v) SDS, 20 $\mu$g/mL Roche Diagnostics RNase A) were added to the resuspended cells. Each Falcon tube was incubated at 37°C for 1 hour. Afterwards, 50 $\mu$L of Roche Diagnostics Proteinase K (at a 20 mg/mL stock concentration) were added. The tubes were gently mixed end-over-end 10 times and then incubated at 50°C for 3 hours. After 1 and 2 hours, all tubes were gently rotated end-over-end. After incubation, 10 mL of buffer saturated phenol (Sigma, phenol solution for molecular biology) were added to each tube, which were placed on a rotator for 15 minutes. The tubes were spun at 2000x*g* for 15 minutes. Then, the aqueous phase was carefully transferred into a clean Falcon tube. We added 5 mL of buffer saturated phenol followed by 5 mL of chloroform. The tubes were rotated again and then spun down for another 15 minutes. The aqueous phase was removed and poured into a new tube. A second chloroform wash was carried out for strains LMECYA269 and LMECYA280. In order for the DNA to precipitate, 4 mL of 5 M ammonium acetate and 30 mL of 100% ice-cold ethanol were added to each tube, which were rotated end-to-end 10 times. Using a sterile plastic hook, we tried to recover as much DNA as possible. However, it proved difficult to recover a significant amount of DNA from most of the tubes. Therefore, we left the 100% ethanol containing tubes at -20°C for 8 hours followed by 1 hour at -80°C. LMECYA269 and LMECYA280 gDNA tubes did not require said freezing step. Then, the tubes were spun down and the pellets were washed twice in 70% ethanol containing Eppendorf tubes. After spinning down at 10000x*g*, the 70% ethanol was removed and the tubes were placed on a heat block for at least 4 hours. One hundred and fifty (150) $\mu$L of 10 mM Tris-HCl (pH 8.5) were added to each tube. All tubes were left for 13 hours at 4°C and then stored at -20°C.

Two R9.5 chemistry flow cells were used for the sequencing step. DNA quality and quantity were assessed using Nanodrop, Qubit and Fragment Analyzer (PROSize 2.0, Advanced Analytical Technologies, Inc.). The six gDNA libraries were prepared using the SQK-RAD003 Rapid Sequencing kit. Reagent and gDNA volumes were adjusted as follows: 15 $\mu$L gDNA, 5 $\mu$L FRA, 1 $\mu$L RPD, 30.5 $\mu$L RBF, 16.5 $\mu$L LLB and 7 $\mu$L NFW. At least 400 ng of gDNA were used for library preparation. Custom MinKNOW (version 1.7.10) scripts allowed us to perform 8-hour sequencing runs with live basecalling and to adjust the initial voltage at the start of each run. The flowcells were washed using EXP-WSH002 and stored for 3 hours between sequencing runs. Moreover, we prepared and sequenced an additional LMECYA167 library using a SQK-RAD002 Rapid Sequencing kit. Library preparation was carried out according to (Jain et al., 2017). Reads were obtained during a 24-hour sequencing run. The genome of LMECYA7 was obtained using both the MinION platform and Illumina MiSeq paired-end sequencing methods. The Illumina dataset was produced by the Laboratório Nacional de Referência das Resistências aos Antimicrobianos e Infeções Associadas aos Cuidados de Saúde (LNR-RA/IACS) along with the Unidade de Tecnologia e Inovação of Instituto Nacional de Saúde Doutor Ricardo Jorge.

### 2.3 Data analysis

Basecalling was performed using Albacore (versions 1.2.1 and 1.2.6, 'read_fast_basecaller.py' script) for all MinION 'pass' reads. Albacore 2.0.2 was also used to rerun basecalling. All fastq files were produced by Albacore. Fasta files were extracted from MinION fast5 data files using Poretools (available at https://poretools.readthedocs.io/en/latest/), version 0.6.0 (Loman and Quinlan, 2014). Moreover, Poretools computes raw read statistics, including minimum, maximum and median length, and produces additional run yield and pore occupancy plots. Read quality versus read length plots were obtained using NanoPlot (available at https://github.com/wdecoster/NanoPlot). Quality score distributions were computed by Poretools. Adapter sequences were removed by Porechop (available at https://github.com/rrwick/Porechop).

Genome assembly was performed for each strain using the Canu assembler (available at https://github.com/marbl/canu) version 1.5 (Koren et al., 2016). We also performed a hybrid LMECYA7 assembly using SPAdes (available at http://bioinf.spbau.ru/spades). A hybrid assembly combines both Illumina and MinION sequencing data to produce a more contiguous genome scaffold. Assembly quality was assessed via QUAST (Gurevich et al., 2013) online tool (http://quast.bioinf.spbau.ru/). Moreover, Nanopolish (available at https://github.com/jts/nanopolish) was used to compute an improved consensus sequence for the LMECYA167 combined data assembly.

Additionally, read alignment was performed using Bowtie 2 (Langmead and Salzberg, 2012), version 2.3.2 (available for download at http://bowtie-bio.sourceforge.net/bowtie2/index.shtml), GraphMap (Sović et al., 2016), available at https://github.com/isovic/graphmap, and NGMLR (Sedlazeck et al., 2017), available at https://github.com/philres/ngmlr. The latter two programs are especially accurate for long third-generation sequencing reads. Lambda phage reads were aligned to the GCA_000840245.1 GenBank/NCBI assembly database reference. LMECYA7 and LMECYA167 were aligned to GCA_000010625.1, while LMECYA269 and LMECYA280 were aligned using the GCA_000710505.1 accession. LEGE06226 and LEGE06233 were aligned to the GCA_000464745.1 scaffold reference. Lastly, we used Samtools (version 1.3.1) to compute aligned read statistics, to obtain the distribution of sequencing depth and to find an overall consensus sequence from some of the aligned data.

We looked for antibiotic resistance genes using the Resistance Gene Identifier (RGI) tool, available at the Comprehensive Antibiotic Resistance Database (CARD) website (https://card.mcmaster.ca/). In order to assess other genomic features present in the obtained genome sequences, we used the Rapid Annotation using Subsystem Technology (RAST) service (available at http://rast.nmpdr.org/). Furthermore, we searched for potential plasmid sequences in our data by running the plasmidspades.py script, which is a part of the SPAdes assembler package.

# 3 Results

## 3.1 First sequencing round

### 3.1.1 Raw read statistics

For the first sequencing round, we started by performing a lambda phage test run followed by the sequencing of two cyanobacterial strain gDNA, LMECYA167 and LEGE06233. We used one R9.4 chemistry flowcell and prepared the library according to the Rapid Sequencing kit standard protocol (SQK-RAD002 kit). Lambda phage sequencing resulted in approximately 207 Mbases in a total of 38,137 reads, whereas for strains LMECYA167 and LEGE06233, we obtained 17,743 and 1,790 reads, respectively. LMECYA167's sequencing run resulted in more than 51 Mbases. For LEGE06233, nearly 4 Mbases were obtained. Moreover, mean read length shows a similar trend to that of total basepair and read yield, that is, lambda phage exhibited much higher values than both cyanobacterial strains. While lambda phage had 5,422 bp mean read length, LMECYA167 and LEGE06233 had 2,912 bp and 2,078 bp, respectively. Similarly, the N50 read statistic was greater for lambda phage (9,342 bp) than for either cyanobacterial strains (approximately 2,000 bp). Other raw read statistics, such as median and minimum read length, can be found in table 2.

Table 2: First sequencing round: Raw read statistics

| Statistics | Lambda | LMECYA167 | LEGE06233 |
|---|---|---|---|
| Total reads | 38137 | 17743 | 1790 |
| Total bases | 206786142 | 51660317 | 3720058 |
| Mean | 5422.19 | 2911.59 | 2078.24 |
| Median | 3346 | 1513 | 1232 |
| Minimum | 5 | 5 | 5 |
| Maximum | 529706 | 602918 | 61302 |
| N25 | 16490 | 13641 | 8000 |
| N50 | 9342 | 4989 | 3249 |
| N75 | 4938 | 2226 | 1724 |

Figure 1 shows read yield over time for all three first sequencing round runs. Whereas read yield was almost linear over time for lambda phage and LMECYA167, LEGE06233 had very poor yield during most of the run time. After 10 hours, LMECYA167 seemed to have reached a plateau during which very few reads were produced.

Quality score per base distribution plots can be found in figure 2. Both lambda phage and LMECYA167 had a 39 maximum base quality score, while LEGE06233 had a maximum of 38. Furthermore, mean quality score values are very different between cyanobacterial and lambda phage read data. Mean quality score is 9.6 and 11.4 for LEGE06233 and LMECYA167, respectively. Contrastingly, mean lambda phage quality score is substantially greater, approximately 17.4. Additionally, read quality score distribution plots reveal one peak around 6 for both cyanobacteria. In the case of lambda phage, there are two quality score peaks: the first around 6 and the second around 13. The highest scoring peak contains most of the lambda sequenced reads. Quality score per read versus read length plots can be found in figures S7, S8 and S9. Overall, lambda phage sequencing had greater base and read yield and produced higher quality reads.

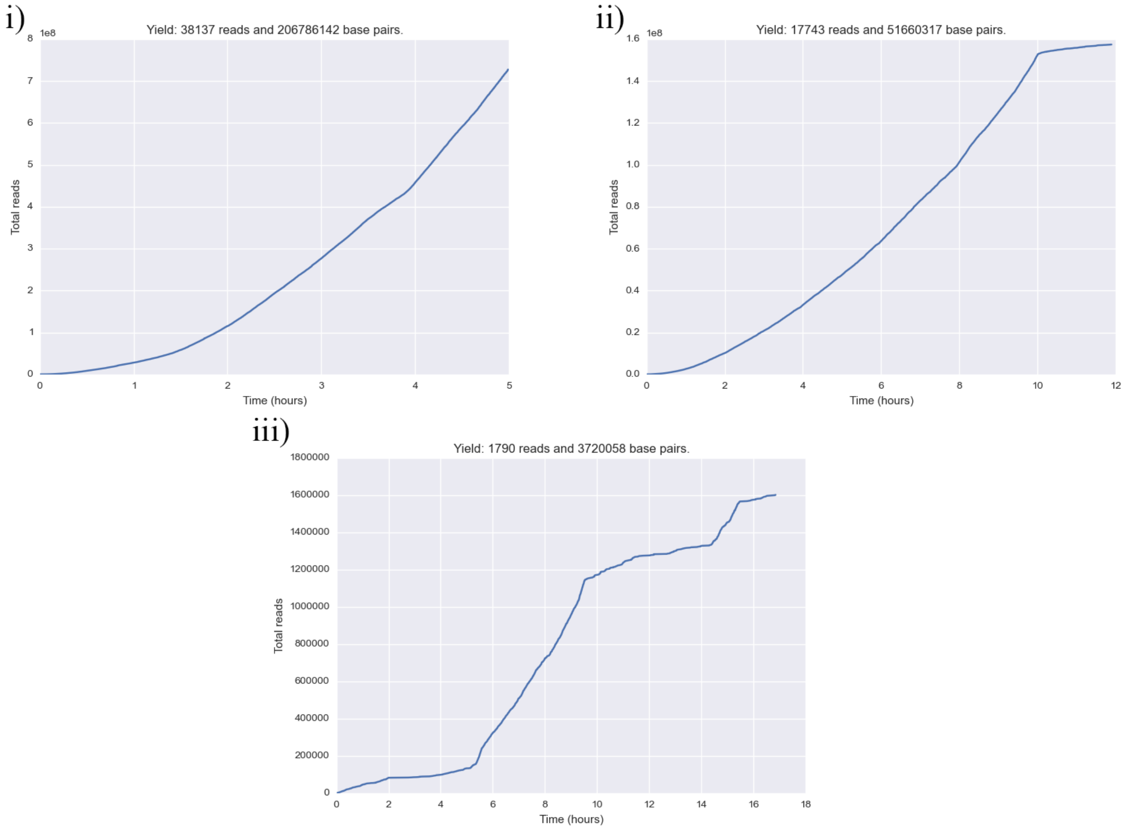

Figure 1: Read yield during the first sequencing round. i) Lambda phage, ii) LMECYA167 and iii) LEGE06233.

After genome assembly and read alignment using Albacore 1.2.1 basecalled reads, we re-ran basecalling with the most recent version of Albacore (version 2.0.2). This improved version applies a quality score filter to the raw data. Only reads with quality scores greater than or equal to 7-8 will pass. In the case of lambda phage data, the total number of reads was reduced by 21% and the total number of sequenced bases by 11%. However, mean read length as well as N50 increased by 12% and 6%, respectively. Maximum read length decreased from 529,706 bp to 61,004 bp. Note that both reads seem to be artifacts, because lambda phage genome length is approximately 48 kbp. This suggests that the quality score filter applied by Albacore 2.0.2 resulted in a greater quality dataset. For LMECYA167 and LEGE06233 strains, each dataset was reduced by nearly 80% and 90% of the total number of reads, respectively. Even though mean read length increased by almost 60% in the case of LMECYA167, maximum read length was reduced by 70%. Moreover, the total number of sequenced bases decreased from 51 Mbases to 16 Mbases. Lastly, regarding LEGE06233, mean read length, total number of sequenced bases and N50 were reduced. All raw read statistics can be found in table 3.

Figure 3 shows base quality score histograms for the three first sequencing round runs. For the lambda phage test run, base quality scores averaged around 20, a 13% increase from the unfiltered data. Moreover, the read quality score plots show a single peak around 12. This observation is in contrast with Albacore 1.2.1 basecalled reads, in which the quality score distribution had two peaks. Furthermore, regarding LMECYA167 and LEGE06233 read quality scores, both score distribution plots exhibit one peak. However, the peak is not around 6 as for the unfiltered data, but around 8-9. This indicates that Albacore 2.0.2 quality score filter is increasing overall dataset quality. For read quality score versus length bivariate plots see figures S10, S11 and S12.

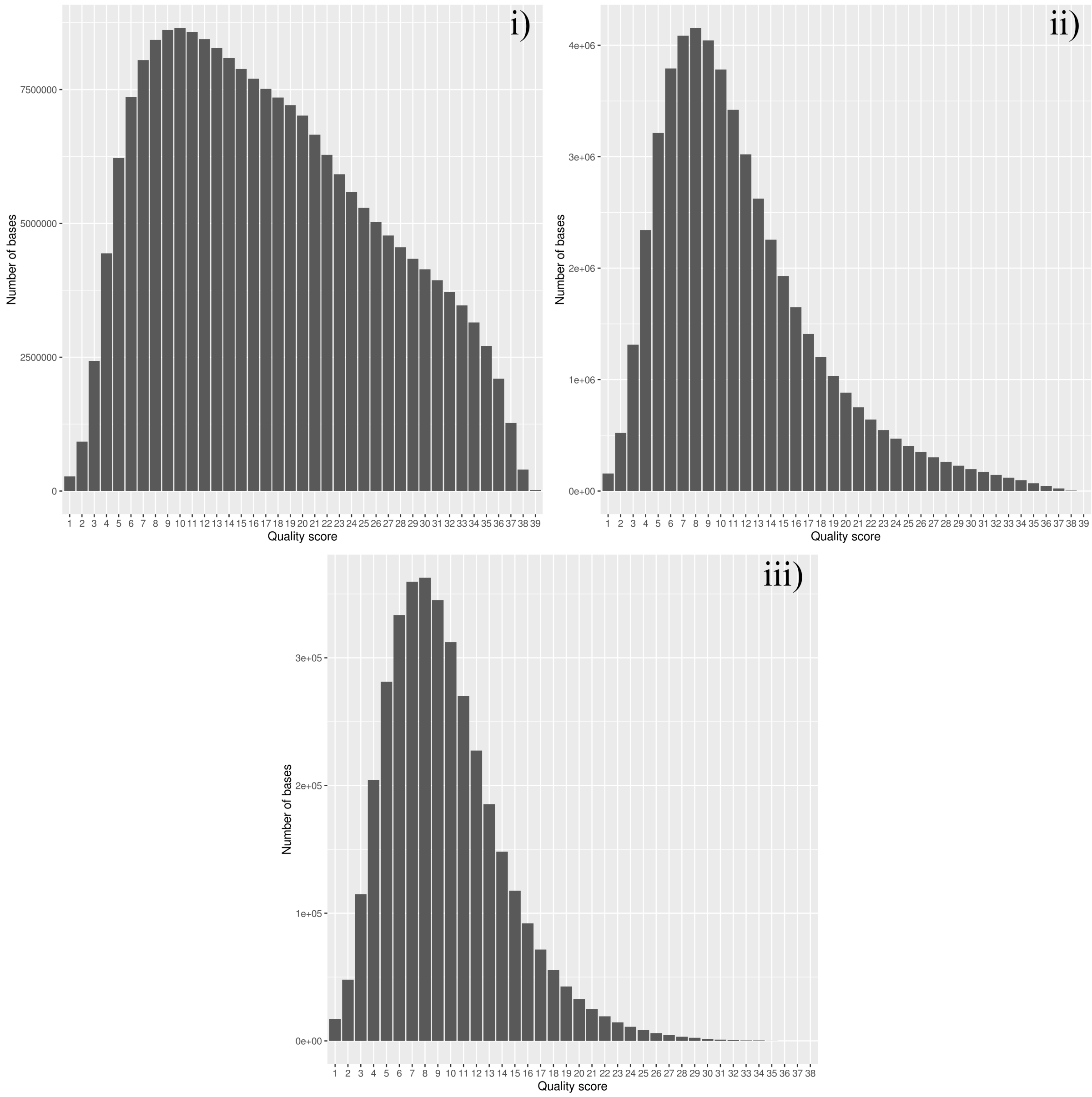

Figure 2: Read quality score distribution for the first sequencing round. i) Lambda phage, ii) LMECYA167 and iii) LEGE06233.

Table 3: First sequencing round: Raw Albacore 2.0.2 basecalled read statistics

| **Statistics** | **Lambda** | **LMECYA167** | **LEGE06233** |
|---|---|---|---|
| Total reads | 29995 | 3550 | 15 |
| Total bases | 182586910 | 16486482 | 29711 |
| Mean | 6087.24 | 4644.08 | 1980.73 |
| Median | 4134 | 3225 | 2109 |
| Minimum | 107 | 50 | 503 |
| Maximum | 61004 | 161357 | 3443 |
| N25 | 16666 | 12201 | 3152 |
| N50 | 9895 | 7098 | 2256 |
| N75 | 5519 | 3957 | 2035 |

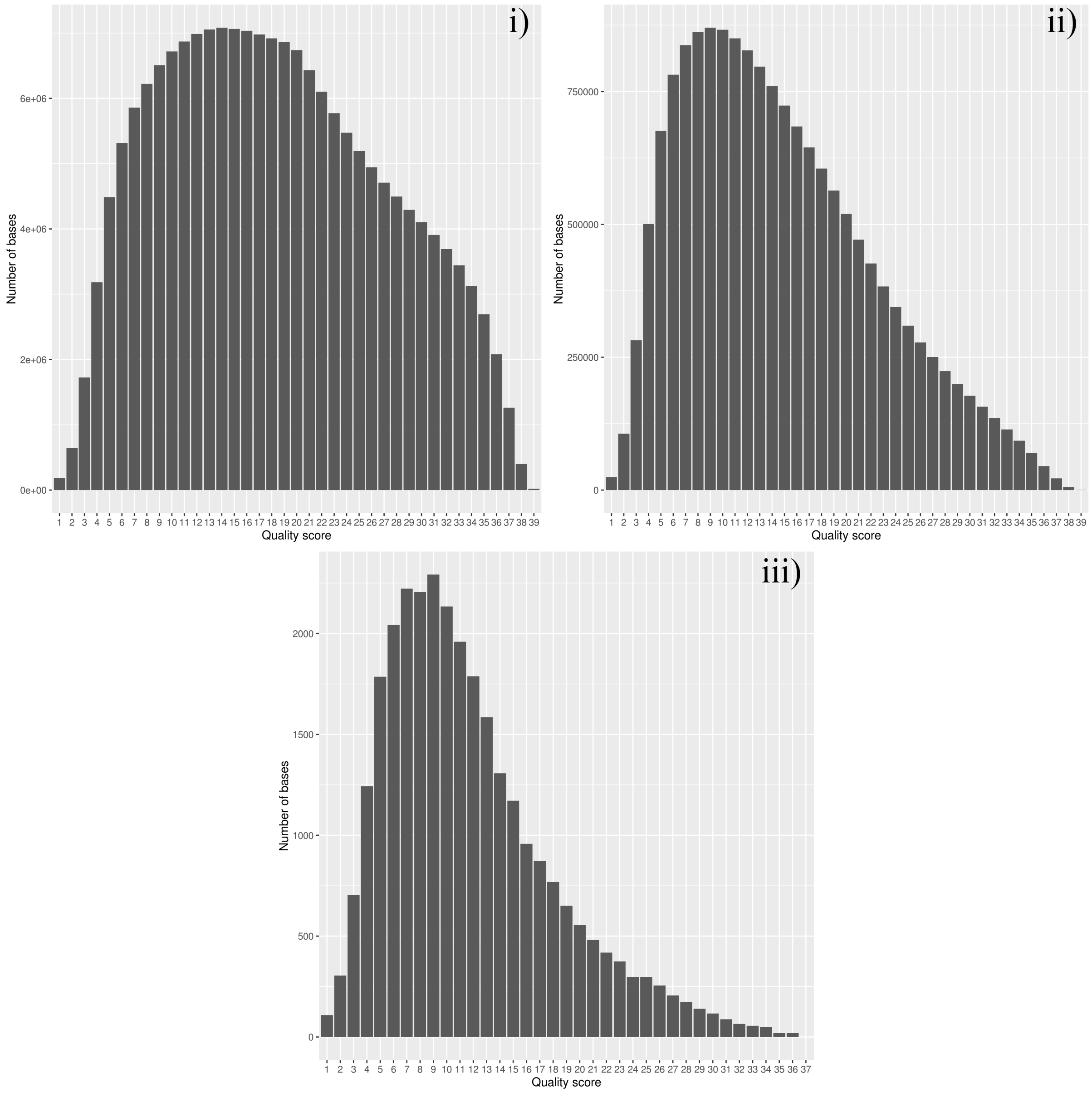

Figure 3: Read quality score distribution for the first sequencing round. i) Lambda phage, ii) LMECYA167 and iii) LEGE06233. Albacore 2.0.2 basecalled reads.

### 3.1.2 Genome assembly

Each independent run was assembled using Canu. Table 4 shows general assembly statistics. For the lambda phage assembly, a single 96,859 bp contig was produced. The contig aligned to the approximately 48 kbp lambda phage genome, resulting in almost full length genome duplication. 8 mismatches and 197 indels were also reported. Overall GC content was 49.9% and its pattern along the genome notably matches that of the reference (figure 4). Furthermore, the assembly produced by LMECYA167 sequencing data resulted in 2 contigs and a total assembly length of 88,606 bp. Neither of the contigs aligned to the reference genome. Lastly, regarding the LEGE06233 sequencing run, no overlaps were found by Canu among the 1,790 reads.

Table 4: First sequencing round: Assembly statistics

| **Statistics** | **Lambda** | **LMECYA167** | **LEGE06233** |
|---|---|---|---|
| No. contigs | 1 | 2 | 0 |
| Total bases | 96859 | 88606 | - |
| Unaligned bases | 0 | 88606 | - |
| Largest contig | 96859 | 79989 | - |
| N50 | 96859 | 79989 | - |
| GC (%) | 49.87 | 45.89 | - |

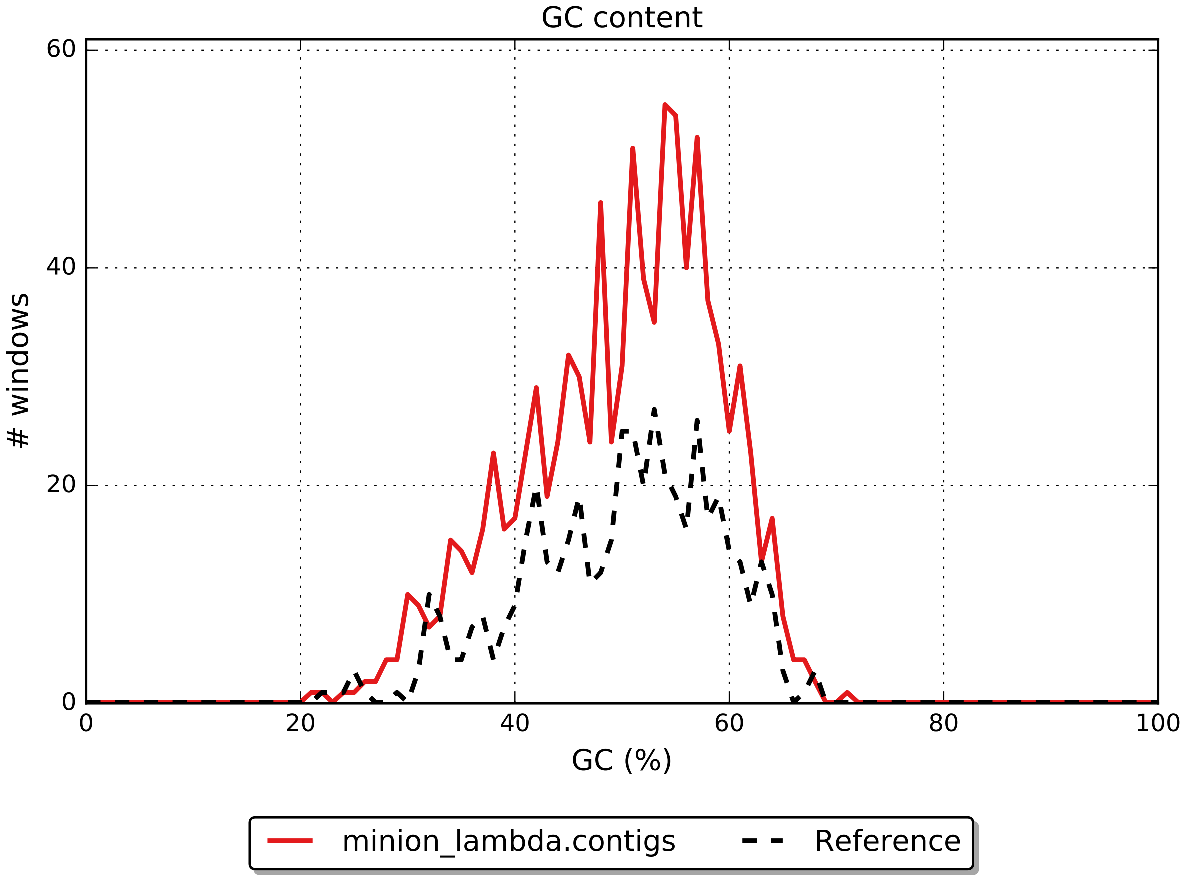

Figure 4: Lambda phage data assembly GC content pattern along the genome. Albacore 1.2.1 basecalled reads.

Genome assembly performed by Canu using Albacore 2.0.2 basecalled reads resulted in one contig for both lambda phage and LMECYA167. In both cases, total assembly length was reduced. The lambda phage assembly revealed a genome duplication ratio of 1.81, since almost the whole 87,881 bp contig aligned to the reference. The assembly had 7 mismatches and 264 indels. GC content pattern along the genome is similar to that of the reference (figure S25). For the LMECYA167 79,639 bp-long scaffold, the contig did not map onto the reference genome. Finally, as with the unfiltered data, the LEGE06233 assembly did not find any overlaps between reads. Assembly statistics can be found in table 5.

Table 5: First sequencing round: Assembly statistics (Albacore 2.0.2 basecalled reads)

| Statistics | Lambda | LMECYA167 | LEGE06233 |
|---|---|---|---|
| No. contigs | 1 | 1 | 0 |
| Total bases | 87881 | 79639 | - |
| Unaligned bases | 0 | 79639 | - |
| Largest contig | 87881 | 79639 | - |
| N50 | 87881 | 79639 | - |
| GC (%) | 49.26 | 50.68 | - |

### 3.1.3 Alignment to reference genome

The Albacore 1.2.1 basecalled reads of the first sequencing round were aligned to a reference using Bowtie2. Results can be found in table 6. The lambda phage dataset generated an alignment from 1,695 reads, i.e. 29.8% of the total number of reads. This assembly resulted in 11,792,034 mapped bp, 1,545,985 mismatches and a 13% error rate. Regarding LMECYA167 and LEGE06233, a negligible amount of reads were aligned to the reference (3 and 8 reads, respectively). Overall, Bowtie2 alignment of cyanobacterial MinION reads was not successful. Sequencing depth plots for the alignment of lambda phage reads onto the reference genome (figure 5) show that most of the depth had a maximum of 2000 times. Moreover, most of the genome alignment had a sequencing depth that was greater than 1000 times.

Table 6: First sequencing round: Bowtie2 alignment statistics

| Statistics | Lambda | LMECYA167 | LEGE06233 |
|---|---|---|---|
| Mapped reads | 1695 (29.8%) | 3 (0.02%) | 8 (0.45%) |
| Mapped bases | 11792034 | 26 | 344 |
| Mismatches | 1545985 | 1 | 1 |
| Error rate | 0.13 | 0.04 | 0.003 |

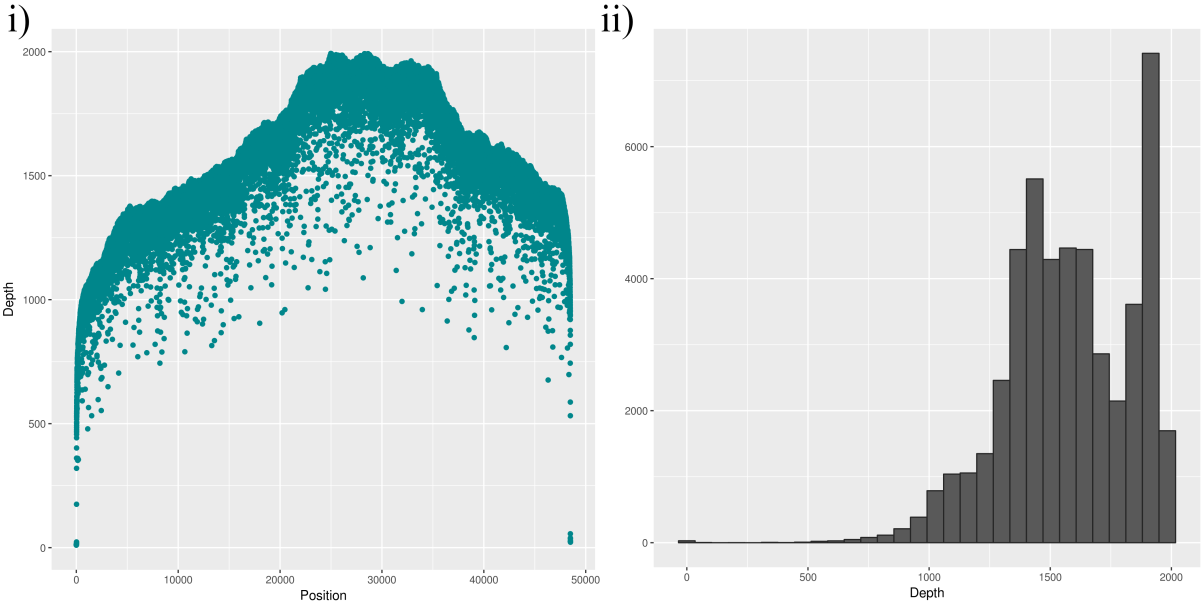

Figure 5: Sequencing depth i) along the genome and ii) overall distribution of lambda phage.

The first sequencing round Albacore 2.0.2 basecalled reads were aligned to the reference using NGMLR, for this software is well suited for MinION dataset alignment. NGMLR alignments were more successful than the previous Bowtie2 runs. For LMECYA167, NGMLR aligned more than 50% of reads resulting in 3,948,374 mapped bases. Although NGMLR increased LEGE06233's overall mapped length from 344 bases to 3573 bases, it corresponds to very little genome coverage. In the case of the lambda phage data, NGMLR aligned nearly 97.67% of Albacore 2.0.2 basecalled reads. This corresponds to 171,279,883 mapped bases, which suggests that the lambda phage alignment was extremely successful with excellent sequencing depth. For both cyanobacterial strains, error rates were high, around 30%, while for lambda phage, the error rate remained low at 13%. These alignment statistics can be found in table 7. Lastly, lambda sequencing depth plots can be found in figure 6. In comparison to Albacore 1.2.1 basecalled reads, Albacore 2.0.2 reads produced an alignment with much higher sequencing depth. Overall depth peaked at over 3000 times with a maximum of 4000 times.

Table 7: First sequencing round: NGMLR alignment statistics of Albacore 2.0.2 basecalled reads

| Statistics | Lambda | LMECYA167 | LEGE06233 |
|---|---|---|---|
| Mapped reads | 29272 (97.59%) | 2188 (50.77%) | 5 (31.25%) |
| Mapped bases | 171279883 | 3948374 | 3573 |
| Mismatches | 21895266 | 11954356 | 1262 |
| Error rate | 0.13 | 0.27 | 0.35 |

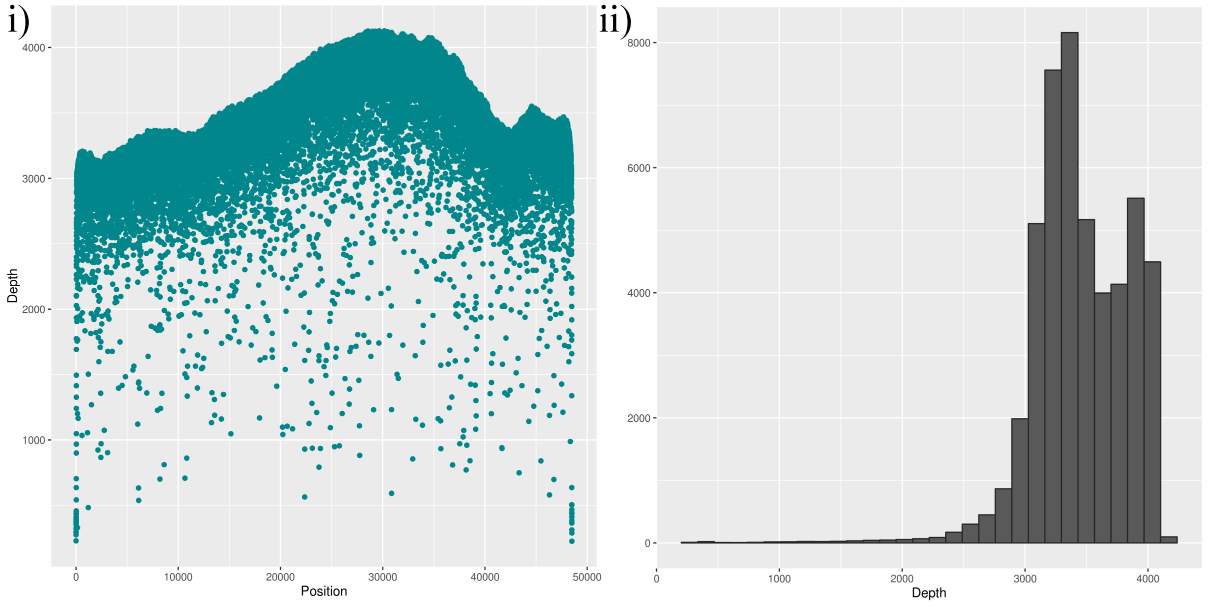

Figure 6: Sequencing depth i) along the genome and ii) overall distribution of lambda phage Albacore 2.0.2 basecalled read alignment.

## 3.2 Second sequencing round

### 3.2.1 Raw read statistics

For the second sequencing round, we performed six 8-hour sequencing runs and an additional 24-hour run. All libraries were prepared using Rapid Sequencing kits. For further details on library preparation see Methods. gDNA quality and quantity assessment can be found in table S1. We used two R9.5

chemistry flowcells to perform said sequencing runs. Each flowcell sequenced well over 100 Mbases and each 8-hour run yielded between 29-66 Mbases. Total number of passed reads varied between 6520, for LMECYA280, and 13110, for LEGE06226. Maximum read length reached 1,586,596 bp for LMECYA167. Moreover, mean read length and N50 were highest for LMECYA167 at 6596 bp and 35977 bp, respectively. Mean read length and N50 were lowest at 3624 bp and 6941 bp for LMECYA7. In the case of the 24-hour LMECYA167 run, 4262 reads were produced with an average read length of 3894 bp. 16,596,861 bases were sequenced during said run. Other raw read statistics are found in table 8.

Table 8: Second sequencing round: Albacore 1.2.6 basecalled read statistics

| Statistics | LMECYA7 | LMECYA167 | LMECYA269 | LMECYA280 | LEGE06226 | LEGE06233 | LMECYA167 24h |
|---|---|---|---|---|---|---|---|
| Total reads | 8950 | 10049 | 9647 | 6520 | 13110 | 12113 | 4262 |
| Total bases | 32434509 | 66286955 | 46789598 | 29591662 | 68710616 | 53713268 | 16596861 |
| Mean | 3623.97 | 6596.37 | 4850.17 | 4538.60 | 5241.08 | 4434.35 | 3894.15 |
| Median | 1721 | 1650 | 2077 | 2219 | 1633 | 1651 | 2281 |
| Minimum | 5 | 5 | 6 | 5 | 5 | 5 | 9 |
| Maximum | 324357 | 1586596 | 307239 | 871568 | 480391 | 777260 | 428391 |
| N25 | 30843 | 88402 | 33490 | 39945 | 52545 | 40072 | 17083 |
| N50 | 6941 | 35977 | 11123 | 8359 | 20130 | 11661 | 5873 |
| N75 | 3031 | 8616 | 4240 | 3709 | 5590 | 4225 | 3025 |

Read yield plots over time (figure 7) reveal an almost linear relationship between said variables during all 8 hour-sequencing runs. Figure 8 shows LMECYA167 read yield during the 24-hour sequencing run. Although such run yielded very few reads over the first 12 hours, afterwards, the relationship between the number of produced reads and run time is almost linear. For pore occupancy over time and read length distribution plots see figures S2 and S3.

Figure 9 shows the histogram of quality score per base for each of the six 8-hour sequencing runs. Average quality scores range from 8.5 (LMECYA7) to 12.4 (LEGE06226). Maximum scores varied between 58 and 59. All strains show similar patterns of quality score distribution, i.e. a single peak around 3-4. This suggests that few good quality reads were produced during said sequencing runs. Read quality score histograms exhibit an identical trend (figures S13 to S18), that is, one peak around 4-5. For some strains, there is a much smaller higher scoring second peak or a wider tail, corresponding to few good quality reads. Quality score per read versus read length plots can be found in figures S13 to S18. Comparing to the first sequencing round quality score statistics, the second round appears to have produced lower quality reads.

All second sequencing round datasets were also basecalled using Albacore 2.0.2. As with the first sequencing round data, the total number of produced reads and sequenced bases were drastically reduced. The number of reads were reduced by at least 73% as was the case for LEGE06233. The greatest reduction of ’pass’ reads was observed for the LMECYA167 24-hour run (98%). Maximum read length was only the same for LMECYA7. The remaining strains had the maximum read length statistic severely reduced. Regarding the N50 read statistic, all strains but LMECYA280 and LEGE06233 revealed an increased N50 value. This suggests higher dataset contiguity in comparison to the unfiltered data. For all raw read statistics, see table 9.

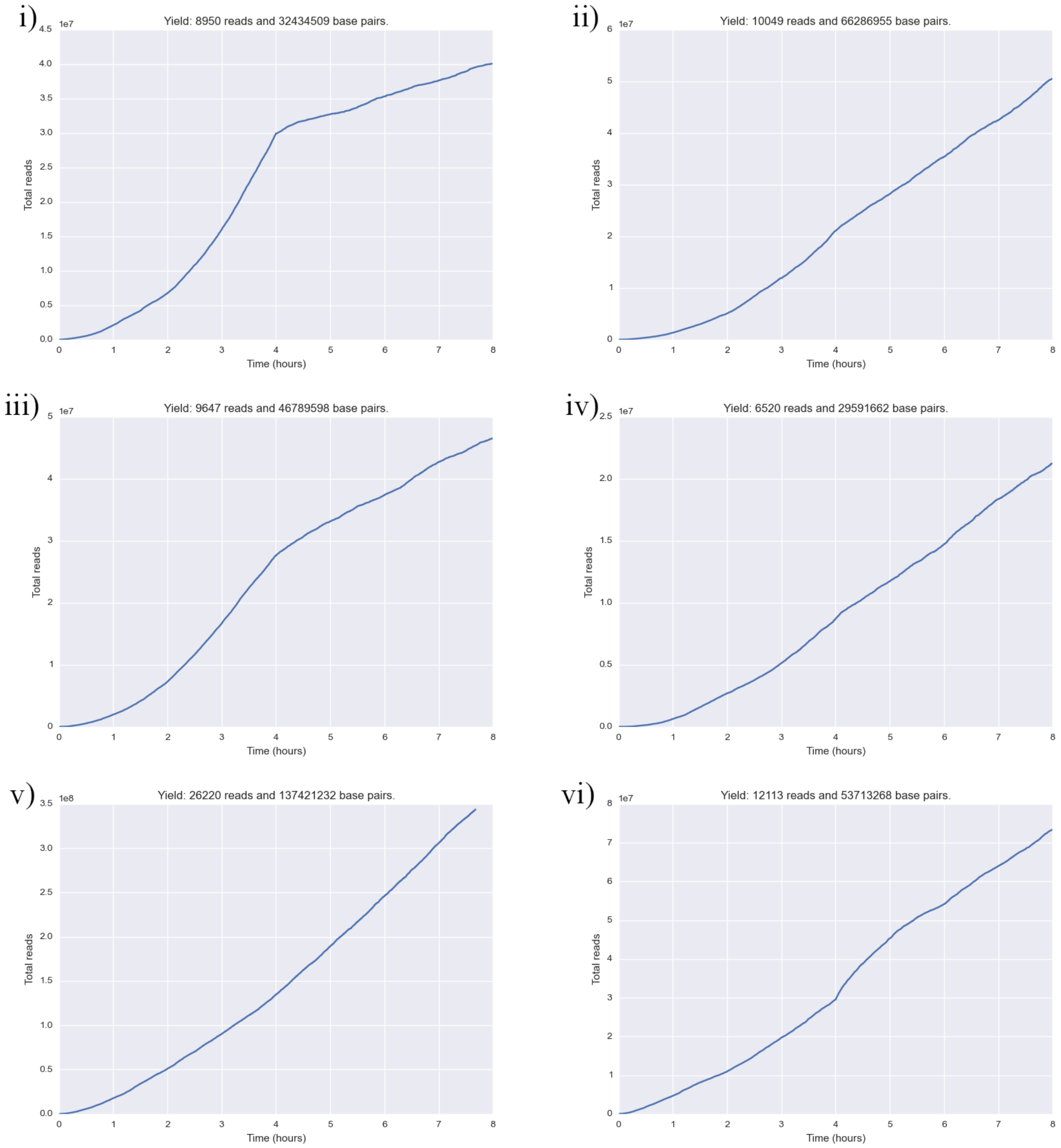

Figure 7: Read yield during the second sequencing round. i) LMECYA7, ii) LMECYA167, iii) LMECYA269, iv) LMECYA280, v) LEGE06226 and vi) LEGE06233.

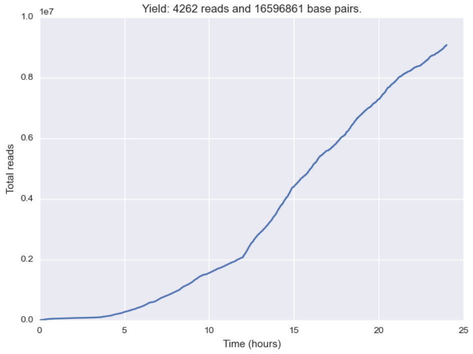

Figure 8: Read yield during the LMECYA167 24h sequencing run.

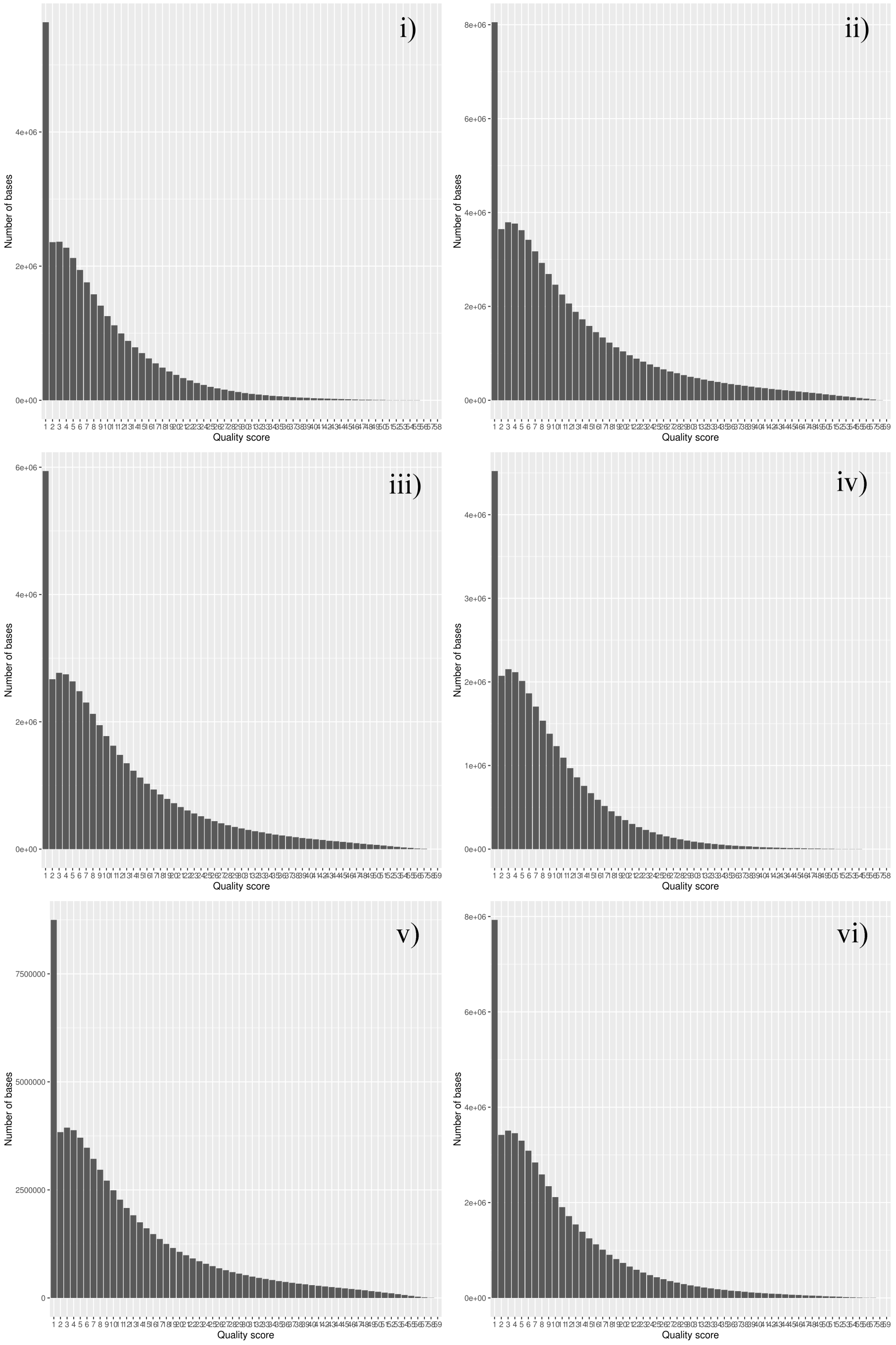

Figure 9: Read quality score distribution for the second sequencing round. i) LMECYA7, ii) LMECYA167, iii) LMECYA269, iv) LMECYA280, v) LEGE06226 and vi) LEGE06233.

The distribution of quality scores per base for each cyanobacterial dataset was also obtained. All histograms show the same single-peak trend as for Albacore 1.2.6 basecalled reads. However, there was a peak shift for all strains but LMECYA7. Each histogram exhibits a quality score peak between 7-9. LMECYA167 appears to have the greatest increase in base quality scores. For quality score per base distributions of Albacore 2.0.2 basecalled reads, see figure 10. Moreover, read quality score distribution plots (figures S19 to S24) reveal a similar increase in peak value. For all strains, quality score peaks vary from 7 to 11.

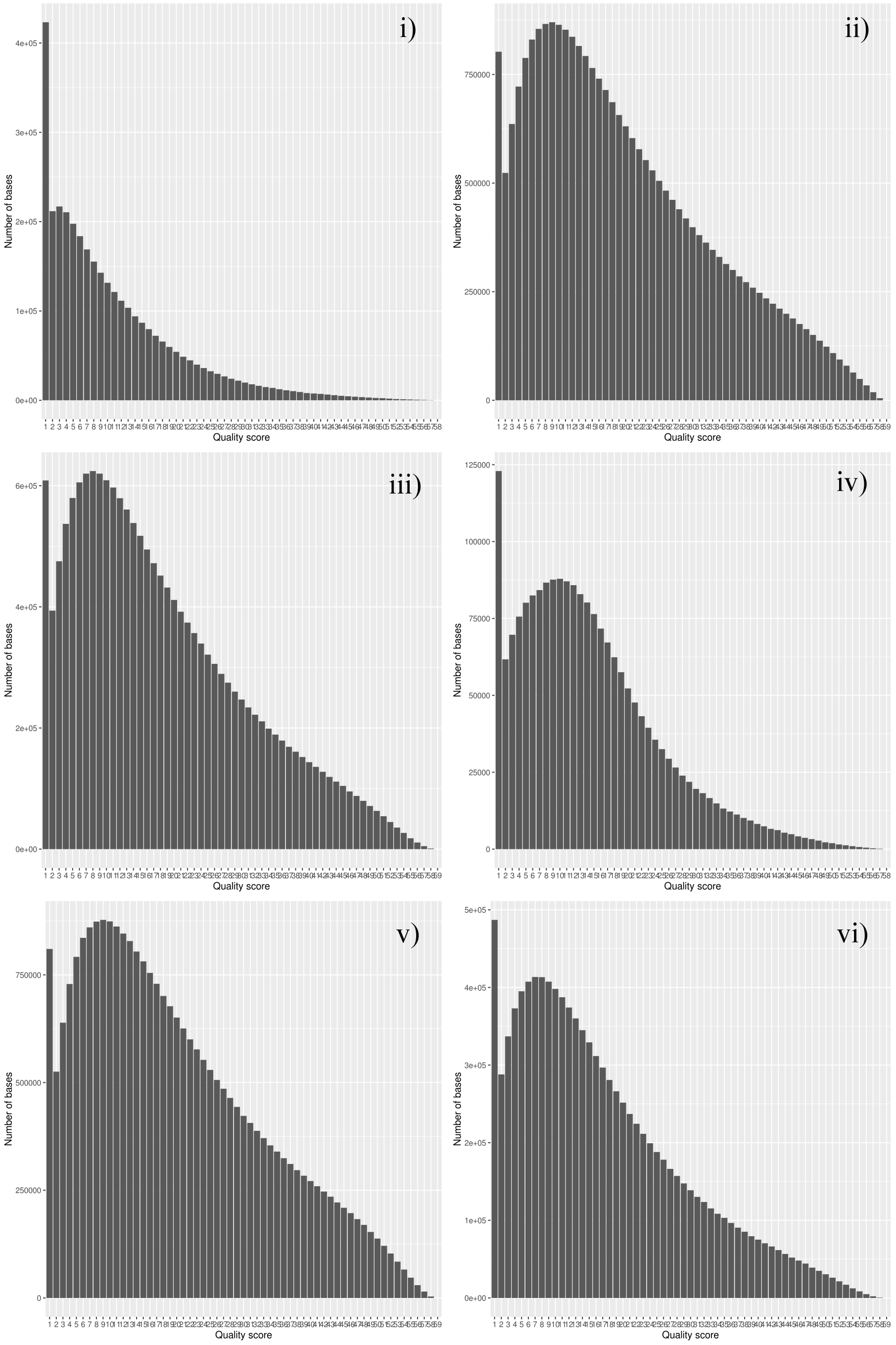

Figure 10: Read quality score distribution for the second sequencing round. i) LMECYA7, ii) LMECYA167, iii) LMECYA269, iv) LMECYA280, v) LEGE06226 and vi) LEGE06233. Albacore 2.0.2 basecalled reads.

Table 9: Second sequencing round: Albacore 2.0.2 basecalled read statistics

| Statistics | LMECYA7 | LMECYA167 | LMECYA269 | LMECYA280 | LEGE06226 | LEGE06233 | LMECYA167 24h |
|---|---|---|---|---|---|---|---|
| Total reads | 424 | 1466 | 2464 | 672 | 2339 | 3217 | 79 |
| Total bases | 3387351 | 25576138 | 16940070 | 2049604 | 26486693 | 10574668 | 730655 |
| Mean | 3387351 | 17446.21 | 6875.03 | 3050.01 | 11323.94 | 3287.12 | 9248.80 |
| Median | 1950 | 8420 | 3339 | 1704 | 5752 | 1387 | 3987 |
| Minimum | 75 | 59 | 35 | 174 | 26 | 58 | 108 |
| Maximum | 324357 | 294791 | 112054 | 295641 | 152016 | 334329 | 71852 |
| N25 | 182110 | 68690 | 30198 | 16381 | 41207 | 16357 | 42610 |
| N50 | 103977 | 40289 | 14348 | 4608 | 23556 | 6957 | 29289 |
| N75 | 8880 | 20373 | 6774 | 2293 | 12393 | 2915 | 8014 |

### 3.2.2 Genome assembly

Table 10 shows general Canu assembly statistics for all seven second sequencing round datasets. While the 8-hour run assemblies resulted in 1-4 contigs, the LMECYA167 24-hour run assembly produced no contigs. Assembled length varied between 1,784 bp for LMECYA167 and 28,403 bp for LMECYA7. Overall, no contig was mapped onto the reference genome.

Table 10: Second sequencing round: Assembly statistics

| Statistics | LMECYA7 | LMECYA167 | LMECYA269 | LMECYA280 | LEGE06226 | LEGE06233 | LMECYA167 24h |
|---|---|---|---|---|---|---|---|
| No. contigs | 1 | 1 | 4 | 2 | - | 2 | 0 |
| Total bases | 28403 | 1784 | 25795 | 9045 | - | 18503 | - |
| Unaligned bases | 28403 | 1784 | 25795 | 9045 | - | 18503 | - |
| Largest contig | 28403 | 1784 | 14628 | 4761 | - | 16870 | - |
| N50 | 28403 | 1784 | 14628 | 4761 | - | 16870 | - |
| GC (%) | 25.24 | 2.41 | 31.76 | 19.49 | - | 30.44 | - |

Albacore 2.0.2 basecalled read assemblies performed by Canu are found in table 11. In contrast to the Albacore 1.2.6 basecalled read assembly, Canu found no overlaps in the LMECYA269 data. Moreover, Canu was able to build a 3805 bp-long assembly for the LMECYA167 24-hour run dataset. LMECYA7, LMECYA280, LEGE06233 and LMECYA167 24-hour run data resulted in a single contig. LEGE06226 assembly did not finish at the time this dissertation was written, despite having run for more than 4 weeks. For most strains, overall assembly length increased. This is particularly notable in the case of LMECYA167 8-hour run data, for which there was an almost 57 times increase in length. Moreover, LMECYA167 assembly was the only one for which a small proportion of the assembled data aligned to the reference (3333 bp). The remaining contigs did not align to the corresponding reference genomes. Even though Albacore 2.0.2 basecalled data resulted in more contiguous assemblies, none of the contigs would have enough length to reach full genome coverage. Lastly, GC content increased for all strains relatively to the Albacore 1.2.6 data, but only LMECYA167 seems to have a similar GC percentage to that of the reference genome (42.7%).

Table 11: Second sequencing round: Assembly statistics (Albacore 2.0.2 basecalled reads)

| Statistics | LMECYA7 | LMECYA167 | LMECYA269 | LMECYA280 | LEGE06226 | LEGE06233 | LMECYA167 24h |
|---|---|---|---|---|---|---|---|
| No. contigs | 1 | 2 | 0 | 1 | 1 | 1 | 1 |
| Total bases | 378042 | 101199 | - | 11825 | 59552 | 5362 | 3805 |
| Unaligned bases | 378042 | 97866 | - | 11825 | 59552 | 5362 | 3805 |
| Largest contig | 378042 | 68809 | - | 11825 | 59552 | 5362 | 3805 |
| N50 | 378042 | 68809 | - | 11825 | 59552 | 5362 | 3805 |
| GC (%) | 48.91 | 40.8 | - | 43.82 | 50.48 | 45.88 | 50.17 |

### 3.2.3 Alignment to reference genome

All second sequencing round datasets were aligned to the corresponding reference genome using two distinct alignment programs, GraphMap and NGMLR. Both software are frequently used to align nanopore data onto reference sequences. GraphMap and NGMLR alignment statistics can be found in tables 12 and 13, respectively. For all datasets, NGMLR was able to increase the proportion of mapped reads at least 6.8 times. LMECYA167 strain showed the greatest increase (over 22 times), having 32.68% of reads mapped by NGMLR versus 4.81% by GraphMap. This increase resulted in a total number of mapped bases well over full genome length, i.e. 7,980,072 mapped bp. None of the remaining alignments had more than 250,000 bp mapped onto the reference. Consequently, none such alignments have nearly enough length for full genome coverage (figures 11, S26 and S27).

Table 12: Second sequencing round: GraphMap alignment statistics

| Statistics | LMECYA7 | LMECYA167 | LMECYA269 | LMECYA280 | LEGE06226 | LEGE06233 | LMECYA167 24h |
|---|---|---|---|---|---|---|---|
| Mapped reads | 45 (0.5%) | 483 (4.81%) | 46 (0.48%) | 31 (0.48%) | 98 (0.75%) | 162 (1.34%) | 23 (0.54%) |
| Mapped bases | 63829 | 6421102 | 187231 | 29777 | 200598 | 174971 | 163546 |
| Mismatches | 24533 | 2797386 | 89642 | 13793 | 86035 | 74866 | 71934 |
| Error rate | 0.38 | 0.44 | 0.48 | 0.46 | 0.43 | 0.43 | 0.44 |

Table 13: Second sequencing round: NGMLR alignment statistics

| Statistics | LMECYA7 | LMECYA167 | LMECYA269 | LMECYA280 | LEGE06226 | LEGE06233 | LMECYA167 24h |
|---|---|---|---|---|---|---|---|
| Mapped reads | 1073 (11.25%) | 4211 (32.68%) | 594 (5.97%) | 351 (5.24%) | 1988 (14.11%) | 1153 (9.13%) | 327 (7.34%) |
| Mapped bases | 89770 | 7980072 | 49589 | 19291 | 237834 | 102594 | 222801 |
| Mismatches | 27262 | 2060977 | 15474 | 6066 | 75045 | 32117 | 63294 |
| Error rate | 0.30 | 0.26 | 0.31 | 0.31 | 0.32 | 0.31 | 0.28 |

After using Albacore 2.0.2 to repeat basecalling, reads were aligned to the reference using NGMLR. We chose NGMLR as it proved to be the most efficient alignment software from our previous analyses. For all strains but the LMECYA167 24-hour run, the total number of mapped bases was reduced by 67% at most, in comparison to the Albacore 1.2.6 basecalled data. The LMECYA167 alignment showed the smallest reduction, i.e. only 2% less bases were mapped onto the reference. Since all datasets were significantly reduced by quality score filtering, every strain had less mapped reads. Nevertheless, LMECYA167's total number of mapped bases was still greater than the reference genome size. Although it might suggest good genome coverage (figure 12), there is a substantial amount of mismatches (1,882,048). On the remaining strain alignments, none could reach significant coverage (figures 12, S28 and S29). Finally, average error rates decreased by almost 3%, from 30% to 27%.

Table 14: Second sequencing round: NGMLR alignment statistics with Albacore 2.0.2 basecalled reads

| **Statistics** | **LMECYA7** | **LMECYA167** | **LMECYA269** | **LMECYA280** | **LEGE06226** | **LEGE06233** | **LMECYA167 24h** |
|---|---|---|---|---|---|---|---|
| Mapped reads | 42 (9.57%) | 2558 (73.80%) | 36 (1.45%) | 78 (10.71%) | 118 (4.92%) | 175 (5.29%) | 46 (54.40%) |
| Mapped bases | 28218 | 7811347 | 18545 | 6335 | 133831 | 47015 | 211418 |
| Mismatches | 6513 | 1882048 | 5392 | 1874 | 39538 | 13431 | 57141 |
| Error rate | 0.23 | 0.24 | 0.29 | 0.30 | 0.29 | 0.29 | 0.27 |

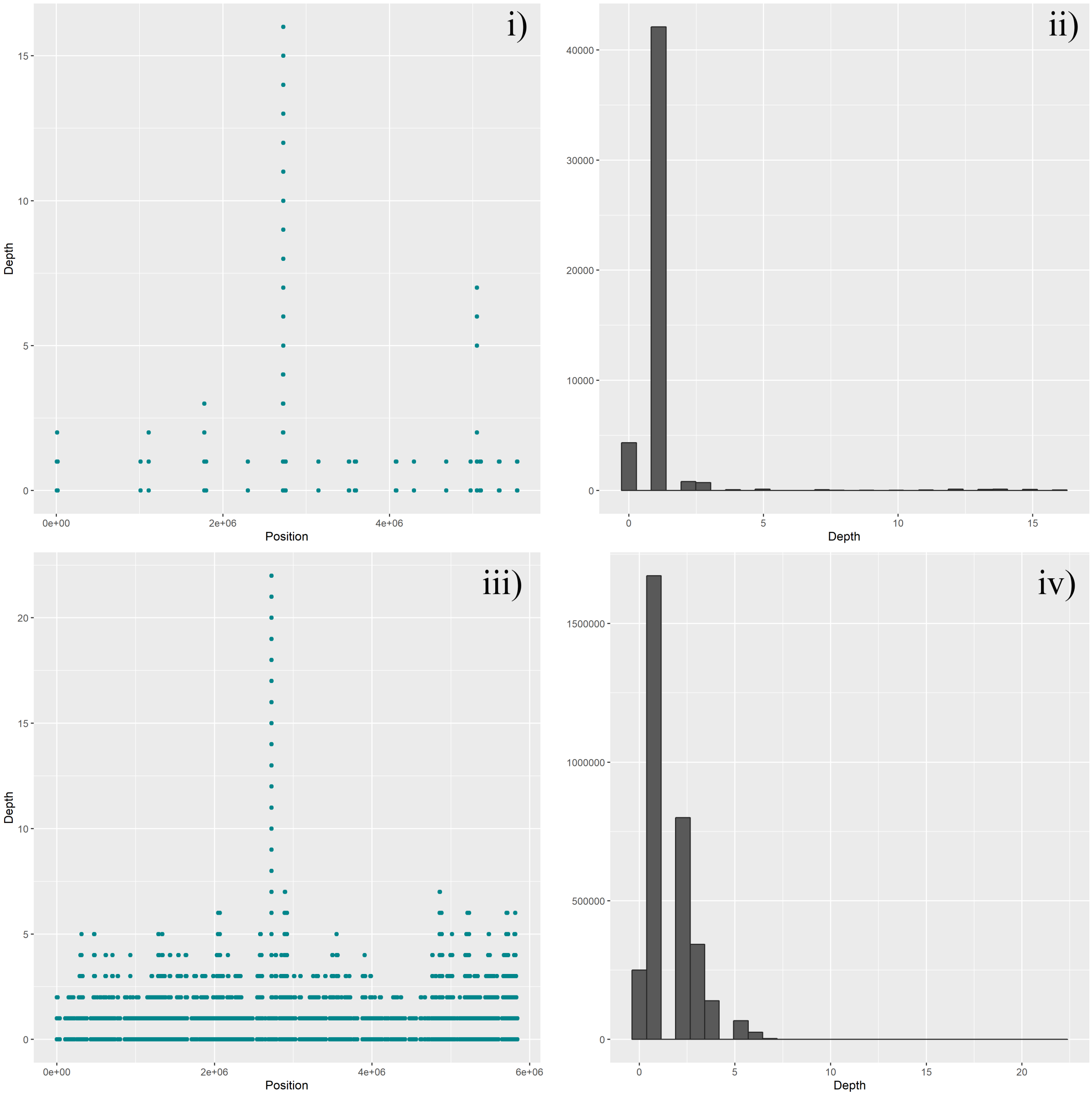

Figure 11: Sequencing depth along the genome (left hand-side plots) and overall distribution (right hand-side plots) of LMECYA7 (i and ii) and LMECYA167 (iii and iv). Alignment performed using Albacore 1.2.6 basecalled reads.

## 3.3 Genome analysis

### 3.3.1 LMECYA7 hybrid assembly

We obtained a hybrid assembly of the genome of LMECYA7 using SPAdes. This *M. aeruginosa* strain assembly resulted from a previously available Illumina paired-end sequencing dataset and our MinION sequencing reads. Two hybrid assemblies were performed for both Albacore 1.2.6 and Albacore 2.0.2 basecalled reads. Table 15 shows general assembly statistics. The total number of contigs of length greater than or equal to 500 bp varies very little between the two assemblies. Accordingly, the total number of assembled bases increased by 1,604 from Albacore 1.2.6 to Albacore 2.0.2 basecalled datasets. Nevertheless, GC content and largest contig length was the same for the two assemblies. Although the total number of assembled bases is well over full genome length, GC content is somewhat different from that of the reference sequence (48.85% vs 42.10%). This indicates that the assembly did not achieve

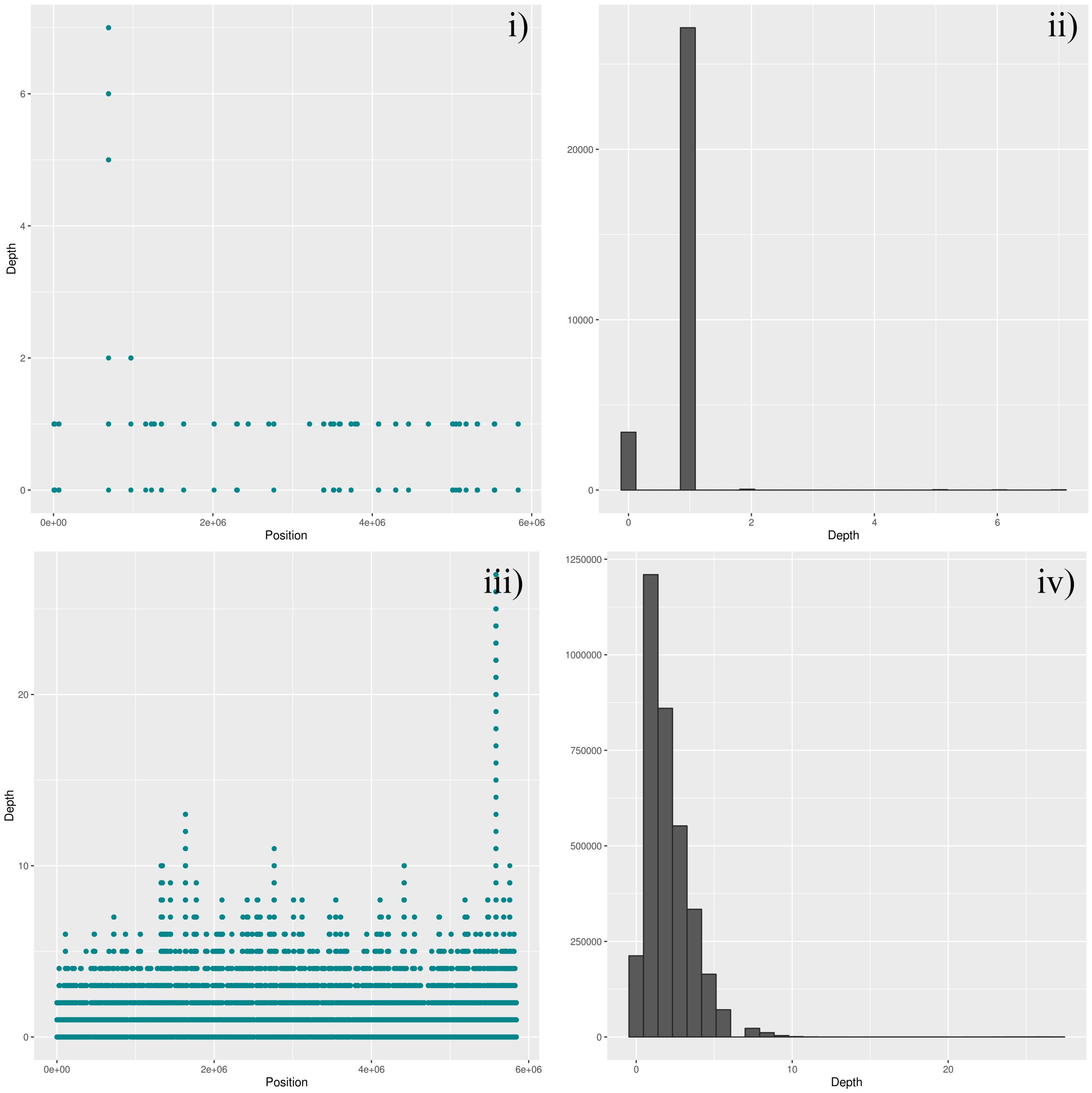

Figure 12: Sequencing depth along the genome (left hand-side plots) and overall distribution (right hand-side plots) of LMECYA7 (i and ii) and LMECYA167 (iii and iv). Alignment performed using Albacore 2.0.2 basecalled reads.

complete genome coverage.

Table 15: LMECYA7 hybrid assembly statistics

| **Statistics** | **LMECYA7 hybrid (Albacore 1.2.6)** | **LMECYA7 hybrid (Albacore 2.0.2)** |
|---|---|---|
| No. contigs | 5674 | 5673 |
| Total bases | 13151544 | 13153148 |
| Largest contig | 380716 | 380716 |
| N50 | 23542 | 23542 |
| GC (%) | 48.85 | 48.85 |

Furthermore, we aligned the two aforementioned hybrid assemblies onto the reference genome. For that purpose, we used three alignment programs: Bowtie2, GraphMap and NGMLR. Although Bowtie2 is a short-read alignment program, the latter two are well suited for longer MinION read alignments. The number of mapped contigs did not vary substantially between the two assemblies. Bowtie2 aligned the least amount of reads (373 and 367 for Albacore 1.2.6 and Albacore 2.0.2 basecalled datasets, respectively) followed by GraphMap. There was a meaningful increase of mapped reads for both NGMLR alignments in comparison to the Bowtie2 and GraphMap. The number of NGMLR mapped bases was 3,807,078 for Albacore 1.2.6 basecalled reads and 3,814,883 for Albacore 2.0.2. This is just over half genome length, indicating that we obtained at least 50% genome coverage from the alignment of the hybrid assembly. For sequencing depth plots, see figure 13.

Table 16: LMECYA7 hybrid assembly alignment statistics

| **Statistics** | **GraphMap** | **Bowtie2** | **NGMLR** |
|---|---|---|---|
| Mapped contigs | 414 (1.74%) | 373 (1.57%) | 1690 (6.74%) |
| Mapped bases | 3064796 | 21596 | 3807078 |
| Mismatches | 1222741 | 273 | 433694 |
| Error rate | 0.40 | 0.01 | 0.11 |

Table 17: LMECYA7 hybrid assembly alignment statistics (Albacore 2.0.2 basecalled reads)

| **Statistics** | **GraphMap** | **Bowtie2** | **NGMLR** |
|---|---|---|---|
| Mapped contigs | 409 (4.16%) | 367 (3.73%) | 1681 (15.06%) |
| Mapped bases | 3232015 | 21246 | 3814883 |
| Mismatches | 1290110 | 269 | 435540 |
| Error rate | 0.40 | 0.01 | 0.11 |

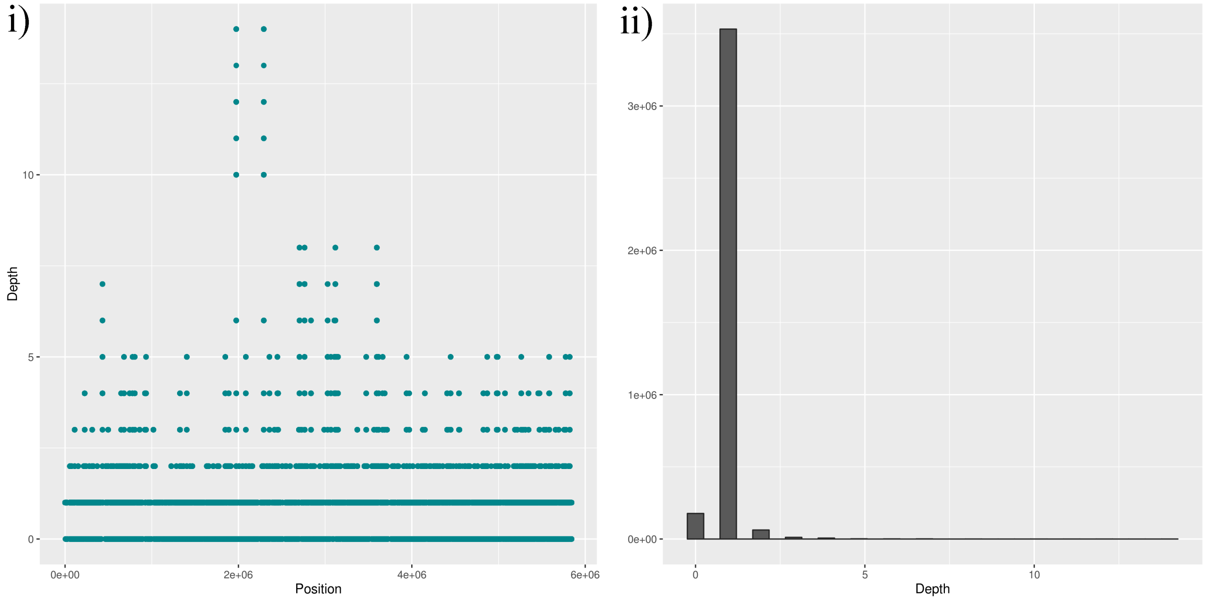

Figure 13: Sequencing depth i) along the genome and ii) overall distribution of LMECYA7 hybrid assembly alignment to the reference genome. Alignment performed using Albacore 2.0.2 basecalled reads.

In summary, we obtained a hybrid assembly from the Albacore 2.0.2 basecalled read dataset and an Illumina paired-end library. The assembled data was aligned onto the reference genome. From the alignment, we built a consensus sequence which was used to look for antibiotic resistance (AR) genes.

We used the Resistance Gene Identifier (RGI) tool available at the Comprehensive Antibiotic Resistance Database (CARD) website to identify all AR genes present in the consensus LMECYA7 sequence. One hundred and forty seven (147) loose AR gene hits (maximum e-value e-10) were found. Loose hits might represent new or more distant homologs of antimicrobial and antibiotic resistance genes. Among such hits, there was evidence of tetracycline resistance genes, such as *tetA*, *tetB* and *tetT*. Moreover, a determinant of sulfonamide resistance, the *sul3* gene, was found. According to the RGI results, this gene product works as an antibiotic target replacement protein. Additionally, 4 beta-lactam resistance determinants were among the RGI loose hits, namely *mecB* and *nmcR*. A DNA-repair protein gene, *mfd*, was also one of the obtained AR gene hits. This protein is associated with fluoroquinolone resistance. Finally, 92 efflux pump gene complexes or subunits that might confer antibiotic resistance were found in the consensus LMECYA7 sequence. For a list of the top 25 hits, see table S2.

Genome annotation revealed 3725 coding sequences. Notably, there is a an alternative folate pathway enzyme that might be responsible for trimethoprim resistance. Since trimethoprim inhibits dihydrofolate reductase (folA) activity, the presence of thymidylate synthase (thyX) might suggest that this strain of *M. aeruginosa* is resistant to trimetrophim due to a folate metabolism alternative enzyme.

### 3.3.2 LMECYA167 combined sequencing data

Lastly, we combined all three LMECYA167 data, i.e. the 12-hour run performed during the first sequencing round, as well as the two second sequencing round runs (8-hour and 24-hour long runs). This resulted in 32,054 Albacore 1.2 basecalled reads and only 5,095 Albacore 2.0.2 basecalled reads.

We performed two Canu assembly runs: one for Albacore 1.2 basecalled reads and another for Albacore 2.0.2 basecalled reads. Assembly statistics can be found in table 18. For the former assembly, 4 contigs resulted in 111,441 bp total length. However, only 23,957 bp mapped onto the reference. Regarding the Albacore 2.0.2 dataset assembly, total assembly length almost doubled, but the overall mapped sequence was only 7867 bp in length. GC content did not vary between the assemblies and it was not substantially different from that of the reference. Lastly, we performed a Nanopolish run on the Albacore 2.0.2 read assembly in order to improve assembly quality and contiguity. Assembly statistics were identical to those of the Albacore 2.0.2 assembly, suggesting that using Nanopolish did not improve assembly quality.

Table 18: LMECYA167 combined sequencing data assembly statistics

| Statistics | Albacore 1.2 | Albacore 2.0.2 | Albacore 2.0.2 + Nanopolish |
|---|---|---|---|
| No. contigs | 4 | 7 | 7 |
| Total bases | 111441 | 200128 | 200128 |
| Unaligned bases | 87484 | 192261 | 192261 |
| Largest contig | 79995 | 79720 | 79720 |
| N50 | 79995 | 41546 | 41546 |
| GC (%) | 44.81 | 44.74 | 44.74 |

Moreover, we used the same alignment software as for the LMECYA7 hybrid assembly alignment. However, Bowtie2 was not able to adequately compute an alignment from the LMECYA167 MinION dataset. Limited memory resources did not allow us to build a LMECYA167 Bowtie2 alignment. Nevertheless, both GraphMap and NGMLR alignments were successful and very similar. For Albacore 1.2 basecalled data, more than 12 Mbases were mapped onto the genome, whereas for Albacore 2.0.2 basecalled reads, overall alignment length was greater than 40 Mbases. Although the number of mapped reads varied greatly between GraphMap and NGMLR, so did the number of mismatches. While GraphMap aligned from 2111 to 2257 reads producing at least 4,871,791 mismatches, NGMLR mapped between 4769 and 7903 reads which resulted in 3,347,028 mismatches at most. Accordingly, error rates were around 41% for GraphMap and 26% for NGMLR. This suggests NGMLR produced the best quality alignments. Sequencing depth and mapping coverage plots for LMECYA167 combined data alignment can be found in figures 14 (Albacore 1.2 basecalled reads), 15 and 16 (Albacore 2.0.2 basecalled reads).

Table 19: LMECYA167 combined sequencing data alignment statistics

| Statistics | GraphMap | Bowtie2 | NGMLR |
|---|---|---|---|
| Mapped reads | 2257 (7.04%) | - | 7903 (21.69%) |
| Mapped bases | 12550924 | - | 12556653 |
| Mismatches | 5149224 | - | 3347028 |
| Error rate | 0.41 | - | 0.27 |

Table 20: LMECYA167 combined sequencing data alignment statistics (Albacore 2.0.2 basecalled reads)

| Statistics | GraphMap | Bowtie2 | NGMLR |
|---|---|---|---|
| Mapped reads | 2111 (41.52%) | - | 4769 (60.73%) |
| Mapped bases | 40357418 | - | 40353861 |
| Mismatches | 4871791 | - | 3031448 |
| Error rate | 0.41 | - | 0.25 |

As with the LMECYA7 assembly alignment, we also obtained a consensus sequence from the aforementioned LMECYA167 assembly. Furthermore, we looked for antibiotic resistance genes by running the RGI tool. We found 147 AR gene loose hits, similarly to the LMECYA7 hybrid assembly. Ninety seven (97) such hits were subunits or complexes of efflux pumps which might confer antibiotic resistance. The sulfonamide resistance gene, *sul3*, was also found. The same tetracycline resistance genes found in LMECYA7 (*tetA*, *tetB* and *tetT*), were present in the LMECYA167 dataset. Regarding beta-lactam resistance, 4 hits were found by RGI. The top 25 antibiotic resistance gene hits can be found in table S3.

The consensus genome sequence was annotated using the RAST service. Four thousand and twenty three (4023) coding sequences were found. Among these, the thymidylate synthase (*thyX*) gene was present, as with LMECYA7. Therefore, trimethoprim resistance mechanism seems to be identical between the two strains.

Finally, we looked for plasmids in the LMECYA167 combined dataset using the plasmidspades.py script. This script is available at the SPAdes assembler website. However, no potential plasmid sequences were found by the script in the dataset.

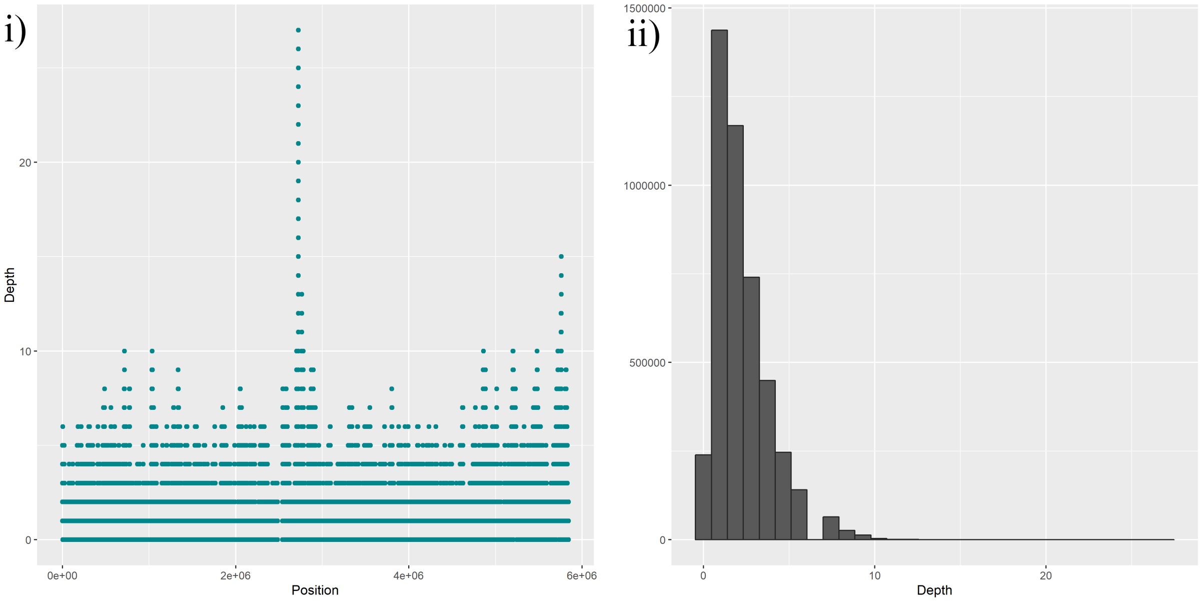

Figure 14: Sequencing depth i) along the genome and ii) overall distribution of LMECYA167 combined data.

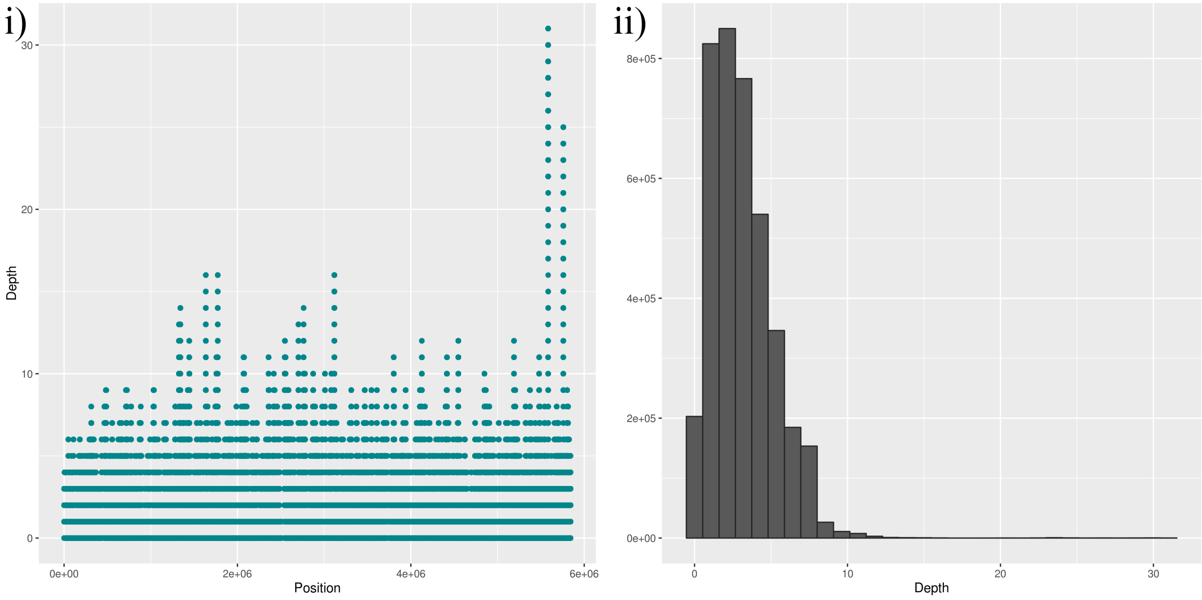

Figure 15: Sequencing depth i) along the genome and ii) overall distribution of LMECYA167 combined data. Alignment performed using Albacore 2.0.2 basecalled reads.

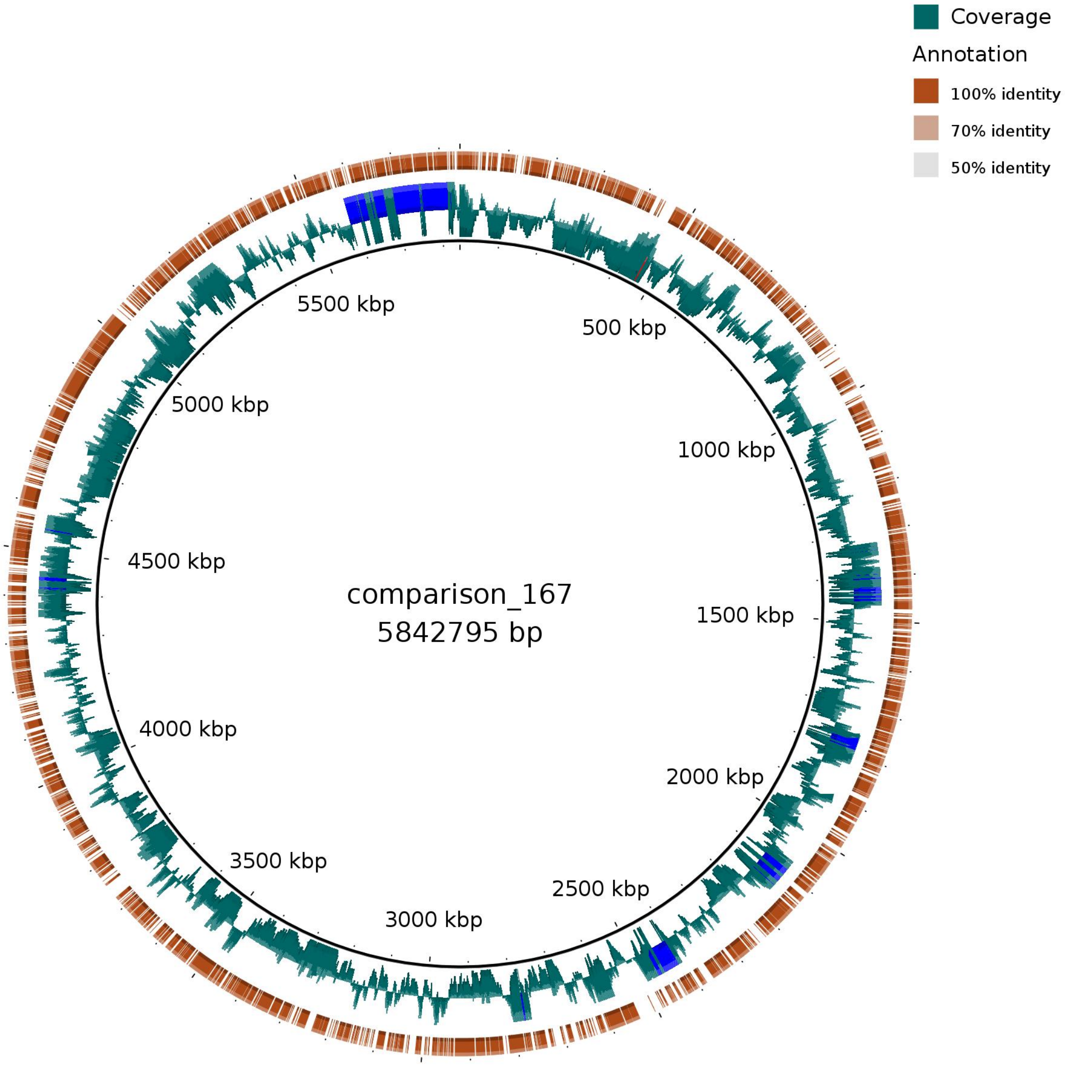

Figure 16: Circular representation of the LMECYA167 combined data scaffold genome. The blue ring corresponds to the mapping coverage along the genome. The orange ring represents all the coding sequences found in LMECYA167 genome scaffold by the RAST annotation tool.

# 4 Discussion

There is evidence of antibiotic resistance genes in the MinION sequenced cyanobacterial gDNA. We obtained genome scaffolds for the two *Microcystis aeruginosa* strains sequenced using MinION. We performed a high molecular weight extraction protocol and aimed for the best quality and quantity gDNA. After quality score filtering, adapter removal, genome assembly and read alignment, we looked for AR genes and other key genomic features in the LMECYA7 and LMECYA167 scaffolds.

Rapid sequencing allowed us to perform several sequencing runs in a matter of days. gDNA library preparation was very simple as well as MinION device usage. Raw read statistics prove that overall read length is much higher that standard high throughput sequencing methods, with reads reaching more than 300,000 bp.

gDNA quality and quantity are the most limiting factors of MinION Rapid Sequencing. Despite all efforts to extract enough DNA to obtain full genome coverage, we did not manage to obtain more than 68% genome sequence. Additionally, most assembly contigs did not map onto the corresponding reference genome. Overall alignment rates were also very low, with few mapped reads. Nevertheless, for both *M. aeruginosa* strains, LMECYA7 and LMECYA167, more than 63% of genes were found.

Furthermore, we have attempted to saturate the flowcell with high molecular weight (HMW) DNA in order to obtain very long reads. Rapid sequencing kits require less gDNA shearing and might allow one to generate such long reads. However, HMW cyanobacterial DNA extraction did not yield as much gDNA as expected. Therefore, MinION sequencing was not as successful as other studies' experiments. Notably, Bayliss et al. (2017) MinION sequenced the 2.86 Mbp-long genome of a *Staphylococcus aureus* strain to a 7-times genomic coverage. Moreover, overall yield per flowcell did not reach the 1 Gbase mark. Other groups generated more than 5 Gbases (Jain et al., 2017).

In spite of numerous difficulties with cyanobacterial DNA, the lambda phage test run seems to have generated good quality and quantity reads. Full genome length coverage was achieved with few mismatches and indels. Albacore 2.0.2 quality score filtering removed far less reads in comparison to the cyanobacterial datasets. This again suggests that input DNA is most crucial for whole-genome sequencing.

Evidence of population adaptation from key genomic features was hindered by the little amount of good quality data acquired. However, in the LMECYA7 and LMECYA167 genome scaffolds, more than 100 homologous antimicrobial resistance genes were found. It suggests that microbial communities have been playing a key role in the evolution of antibiotic resistance phenotypes of these strains of cyanobacteria. In particular, the presence of the thymidylate synthase (*thyX*) gene in both *M. aeruginosa* strains indicates that this lineage evolved to become resistant to trimethoprim through the acquisition of an alternative enzyme in the folate pathway, as suggested by Dias et al. (2015) for cyanobacteria and Myllykallio et al. (2003) for other bacteria. The antibiotic resistance mechanism seems to be identical in both *M. aeruginosa* strains. Moreover, LMECYA7 and LMECYA167's nalidixic acid resistance might be explained by the *mfd* gene. This gene's product is a DNA-repair protein involved in fluoroquinolone resistance phenotypes, according to the CARD database results. Although LMECYA7 and LMECYA167 do not appear to be resistant to amoxicillin (Dias et al., 2015), a few beta-lactam resistance associated genes were found by the RGI tool. Lastly, tetracycline resistance seems be associated with the presence of several *tetA*, *tetB* and *tetT* genes in these strains' genomes.

In order to fully grasp the evolution of cyanobacterial lineages as their genomes are shaped by microbial community pressures, a more comprehensive number of samples has to be sequenced. It would been interesting to make use of the Estela Sousa e Silva Algae Culture Collection (ESSACC) of cyanobacterial isolates across Portugal and sequence more genomes. We would then be able to characterise the resistome of cyanobacterial communities in Portugal. Investigating this pool of AR genes would enable us to identify horizontal gene transfer events as well as point mutations that confer antibiotic resistance. Furthermore, it would be of interest to collect samples along a given time period and perform a time-series study of cyanobacterial communities. Sequencing these individuals' genomes would allow us to track allele frequency shifts that could be associated with fluctuating natural selection.

Similarly to MinION metagenomic studies, such as Schmidt et al. (2017) and Brown et al. (2017), it would be interesting to combine all cyanobacterial read data and attempt at species-level identification. If successful identification was achieved, we would be able to discard gDNA extraction contamination as a source of poor quality reads.

Regarding MinION sequencing of cyanobacterial gDNA, future sequencing rounds would require further protocol optimisation. On the HMW extraction protocol, overall yield and contaminant removal would have to be improved. DNA quantity was not enough to obtain full genome coverage and additional purification steps would also be required. Furthermore, ONT's Rapid Sequencing kits are not ideal if one is aimed at complete genome sequencing. The 1D$^2$ Sequencing Kit provides the user with higher accuracy reads and it requires less input DNA. Finally, increasing run time from 8-hour sequencing protocols would allow one to collect more data that could improve both depth and coverage. LMECYA167 combined data is to show that a total sequencing run time of 44 hours increased overall yield that contributed to almost 68% coverage.

In summary, MinION Rapid Sequencing has proved to be an easy-to-use sequencing method that allows for real-time monitoring of each run. To obtain maximum MinION read quality and yield, it is necessary to optimise the process of gDNA extraction and purification. Despite some underwhelming results, MinION sequencing seems to produce good quality data if the best input DNA is provided. Given that, sequencing more cyanobacterial genomes will help to further characterise the resistome of microbial communities of freshwater bodies in Portugal. Moreover, it will allow us to track the evolution of antibiotic resistance phenotypes and thus collect evidence of population adaptation through the acquisition of AR genes.

# 5 References

Altermann, Wladyslaw and Józef Kazmierczak (2003). “Archean microfossils: a reappraisal of early life on Earth”. In: *Research in Microbiology* 154.9, pp. 611–617.

Amos, G.C.A., L Zhang, P.M. Hawkey, W.H. Gaze, and E.M. Wellington (2014). “Functional metagenomic analysis reveals rivers are a reservoir for diverse antibiotic resistance genes”. In: *Veterinary Microbiology* 171.3-4, pp. 441–447.

Ashton, Philip M, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O’Grady (2014). “MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island”. In: *Nature Biotechnology* 33.3, pp. 296–300.

Bayliss, Sion C., Vicky L. Hunt, Maho Yokoyama, Harry A. Thorpe, and Edward J. Feil (2017). “The use of Oxford Nanopore native barcoding for complete genome assembly”. In: *GigaScience* February, pp. 1–6.

Benítez-Páez, Alfonso and Yolanda Sanz (2017). “Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION portable nanopore sequencer”. In: *GigaScience* 6.7, pp. 1–12.

Bennett, Albert F and Richard E Lenski (2007). “An experimental test of evolutionary trade-offs during temperature adaptation.” In: *Proceedings of the National Academy of Sciences of the United States of America* 104 Suppl, pp. 8649–8654.

Bennett, P M (2009). “Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria”. In: *British Journal of Pharmacology* 153.S1, S347–S357.

Briand, Enora, Claude Yéprémian, Jean François Humbert, and Catherine Quiblier (2008). “Competition between microcystin- and non-microcystin-producing Planktothrix agardhii (cyanobacteria) strains under different environmental conditions”. In: *Environmental Microbiology* 10.12, pp. 3337–3348.

Brown, Bonnie L., Mick Watson, Samuel S. Minot, Maria C. Rivera, and Rima B. Franklin (2017). “MinION nanopore sequencing of environmental metagenomes: a synthetic approach”. In: *GigaScience* 6.3, pp. 1–10.

Calteau, Alexandra, David P Fewer, Amel Latifi, Thérèse Coursin, Thierry Laurent, Jouni Jokela, Cheryl A Kerfeld, Kaarina Sivonen, Jörn Piel, and Muriel Gugger (2014). “Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria.” In: *BMC Genomics* 15.1, p. 977.

Chen, You, C. Kay Holtman, Roy D. Magnuson, Philip A. Youderian, and Susan S. Golden (2008). “The complete sequence and functional analysis of pANL, the large plasmid of the unicellular freshwater cyanobacterium Synechococcus elongatus PCC 7942”. In: *Plasmid* 59.3, pp. 176–192.

Christiansen, Guntram, Carole Molitor, Benjamin Philmus, and Rainer Kurmayer (2008). “Nontoxic strains of cyanobacteria are the result of major gene deletion events induced by a transposable element”. In: *Molecular Biology and Evolution* 25.8, pp. 1695–1704.

Christiansen, Guntram, Alexander Goesmann, and Rainer Kurmayer (2014). “Elucidation of insertion elements carried on plasmids and in vitro construction of shuttle vectors from the toxic cyanobacterium Planktothrix”. In: *Applied and Environmental Microbiology* 80.16, pp. 4887–4897.

Davies, J (1994). “Inactivation of antibiotics and the dissemination of resistance genes”. In: *Science* 264.5157, pp. 375–382.

Dias, Elsa, Micaela Oliveira, Daniela Jones-Dias, Vitor Vasconcelos, Eugénia Ferreira, Vera Manageiro, and Manuela Caniça (2015). “Assessing the antibiotic susceptibility of freshwater cyanobacteria spp.” In: *Frontiers in Microbiology* 6.JUL, pp. 1–11.

Edwards, Arwyn, Aliyah R Debbonaire, Birgit Sattler, Luis AJ Mur, and Andrew J Hodson (2016). "Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N". In: *bioRxiv*, p. 073965.

Eriksson, J E, L Gronberg, S Nygard, J P Slotte, and J A O Meriluoto (1990). "Hepatocellular Uptake of H-3 Dihydromicrocystin-Lr, a Cyclic Peptide Toxin". In: *Biochimica Et Biophysica Acta* 1025.1, pp. 60–66.

Felsenstein, Joseph (1974). "The evolutionary advantage of recombination. II. Individual selection for recombination". In: *Genetics* 78.2, pp. 737–756.

Fisher, Ronald Aylmer (1930). *The genetical theory of natural selection*. Oxford University Press, London, p. 272.

Francis, George (1878). "Poisonous Australian lake". In: *Nature* 18.444, pp. 11–12.

Giordano, Francesca, Louise Aigrain, Michael A Quail, Paul Coupland, James K Bonfield, Robert M Davies, German Tischler, David K Jackson, Thomas M Keane, Jing Li, Jia-Xing Yue, Gianni Liti, Richard Durbin, and Zemin Ning (2017). "De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms." In: *Scientific Reports* 7.1, p. 3935.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler (2013). "QUAST: Quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8, pp. 1072–1075.

Hess, Wolfgang R. (2011). "Cyanobacterial genomics for ecology and biotechnology". In: *Current Opinion in Microbiology* 14.5, pp. 608–614.

Hesse, Karina, Elke Dittmann, and Thomas Börner (2001). "Consequences of impaired microcystin production for light-dependent growth and pigmentation of Microcystis aeruginosa PCC 7806". In: *FEMS Microbiology Ecology* 37.1, pp. 39–43.

Hoekstra, Hopi E, Rachel J Hirschmann, Richard A Bundey, Paul A Insel, and Janet P Crossland (2006). "A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern". In: *Science* 313.5783, pp. 101–104.

Huddleston, Jennifer R. (2014). "Horizontal gene transfer in the human gastrointestinal tract: Potential spread of antibiotic resistance genes". In: *Infection and Drug Resistance* 7, pp. 167–176.

Humbert, Jean François, Valérie Barbe, Amel Latifi, Muriel Gugger, Alexandra Calteau, Therese Coursin, Aurélie Lajus, Vanina Castelli, Sophie Oztas, Gaëlle Samson, Cyrille Longin, Claudine Medigue, and Nicole Tandeau de Marsac (2013). "A Tribute to Disorder in the Genome of the Bloom-Forming Freshwater Cyanobacterium Microcystis aeruginosa". In: *PLoS ONE* 8.8.

Istace, Benjamin, Anne Friedrich, Léo D'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti, Arnaud Lemainque, Stefan Engelen, Patrick Wincker, Joseph Schacherer, and Jean-Marc Aury (2016). "de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer". In: *BioRxiv*, p. 066613.

Jain, M, S Koren, J Quick, A C Rand, T A Sasani, J R Tyson, A D Beggs, A T Dilthey, I T Fiddes, S Malla, H Marriott, K H Miga, T Nieto, J O'Grady, H E Olsen, B S Pedersen, A Rhie, H Richardson, A R Quinlan, T P Snutch, L Tee, B Paten, A M Phillippy, J T Simpson, N J Loman, and M Loose (2017). "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *BioRxiv*. arXiv: `128835`.

Jansen, Hans J., Michael Liem, Suzanne A. Jong-Raadsen, Sylvie Dufour, Finn-Arne Weltzien, William Swinkels, Alex Koelewijn, Arjan P. Palstra, Bernd Pelster, Herman P. Spaink, Guido E. Van den Thillart, Ron P. Dirks, and Christiaan V. Henkel (2017). "Rapid de novo assembly of the European eel genome from nanopore sequencing reads". In: *BioRxiv*.

Jiao, Ya-Nan, Hong Chen, Rui-Xia Gao, Yong-Guan Zhu, and Christopher Rensing (2017). *Organic compounds stimulate horizontal transfer of antibiotic resistance genes in mixed wastewater treatment systems*. Vol. 184. Elsevier Ltd, pp. 53–61.

Jungblut, Anne Dorothee and Brett A. Neilan (2006). “Molecular identification and evolution of the cyclic peptide hepatotoxins, microcystin and nodularin, synthetase genes in three orders of cyanobacteria”. In: *Archives of Microbiology* 185.2, pp. 107–114.

Kaebernick, Melanie and Brett A. Neilan (2001). “Ecological and molecular investigations of cyanotoxin production”. In: *FEMS Microbiology Ecology* 35.1, pp. 1–9.

Karl, D., A. Michaels, B. Bergman, Douglas G. Capone, Edward J. Carpenter, R. Letelier, F. Lipschultz, H. Paerl, D. Sigman, and L. Stal (2002). “Dinitrogen fixation in the world’s oceans”. In: *Biogeochemistry* 57-58, pp. 47–98.

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy (2016). “Canu : scalable and accurate long- read assembly via adaptive k - mer weighting and repeat separation”. In: pp. 1–35. arXiv: 071282.

Kurmayer, Rainer and Marlies Gumpenberger (2006). “Diversity of microcystin genotypes among populations of the filamentous cyanobacteria Planktothrix rubescens and Planktothrix agardhii”. In: *Molecular Ecology* 15.12, pp. 3849–3861.

Kurmayer, Rainer, Guntram Christiansen, Marlies Gumpenberger, and Jutta Fastner (2005). “Genetic identification of microcystin ecotypes in toxic cyanobacteria of the genus Planktothrix”. In: *Microbiology* 151.5, pp. 1525–1533.

Kurmayer, Rainer, Judith F Blom, Li Deng, and Jakob Pernthaler (2015). “Integrating phylogeny, geographic niche partitioning and secondary metabolite synthesis in bloom-forming Planktothrix.” In: *The ISME journal* 9.4, pp. 909–21.

Langmead, Ben and Steven L Salzberg (2012). “Fast gapped-read alignment with Bowtie 2.” In: *Nature methods* 9.4, pp. 357–9. arXiv: {\#}14603.

Laver, T., J. Harrison, P. A. O’Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme (2015). “Assessing the performance of the Oxford Nanopore Technologies MinION”. In: *Biomolecular Detection and Quantification* 3, pp. 1–8.

Loman, Nicholas J. and Aaron R. Quinlan (2014). “Poretools: A toolkit for analyzing nanopore sequence data”. In: *Bioinformatics* 30.23, pp. 3399–3401.

Lu, Hengyun, Francesca Giordano, and Zemin Ning (2016). “Oxford Nanopore MinION Sequencing and Genome Assembly”. In: *Genomics, Proteomics and Bioinformatics* 14.5, pp. 265–279.

Ludden, Catherine, Sandra Reuter, Kim Judge, Theodore Gouliouris, Beth Blane, Francesc Coll, Plamena Naydenova, Martin Hunt, Alan Tracey, Katie L. Hopkins, Nicholas M. Brown, Neil Woodford, Julian Parkhill, and Sharon J. Peacock (2017). “Sharing of carbapenemase-encoding plasmids between Enterobacteriaceae in UK sewage uncovered by MinION sequencing”. In: *Microbial Genomics* 3.7.

Manageiro, Vera, Eugénia Ferreira, Manuela Caniça, and Célia M. Manaia (2014). “GES-5 among the beta-lactamases detected in ubiquitous bacteria isolated from aquatic environment samples”. In: *FEMS Microbiology Letters* 351.1, pp. 64–69.

Marti, Elisabet, Juan Jofre, and Jose Luis Balcazar (2013). “Prevalence of Antibiotic Resistance Genes and Bacterial Community Composition in a River Influenced by a Wastewater Treatment Plant”. In: *PLoS ONE* 8.10, pp. 1–8.

Martins, Joana, Luísa Peixe, and Vítor Vasconcelos (2010). “Cyanobacteria and bacteria co-occurrence in a wastewater treatment plant: Absence of allelopathic effects”. In: *Water Science and Technology* 62.8, pp. 1954–1962.

Michael, Todd P, Florian Jupe, Felix Bemm, Timothy Motley, Justin P Sandoval, Olivier Loudet, Detlef Weigel, and Joseph R Ecker (2017). “High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell”. In: *BioRxiv*, pp. 1–18.

Mundy, Nicholas I (2005). “A window on the genetics of evolution: MC1R and plumage colouration in birds”. In: *Proceedings of the Royal Society B: Biological Sciences* 272.1573, pp. 1633–1640.

Myllykallio, Hannu, Damien Leduc, Jonathan Filee, and Ursula Liebl (2003). “Life without dihydrofolate reductase FolA”. In: *Trends in Microbiology* 11.5, pp. 220–223.

Paulino, Sérgio, Filomena Sam-Bento, Catarina Churro, Elsa Alverca, Elsa Dias, Elisabete Valério, and Paulo Pereira (2009). “The estela sousa e silva algal culture collection: A resource of biological and toxicological interest”. In: *Hydrobiologia* 636.1, pp. 489–492.

Prasanna, R., K. Madhan, R. N. Singh, A. K. Chauhan, and L. Nain (2010). “Developing Biochemical and Molecular Markers for Cyanobacterial Inoculants”. In: *Folia Microbiologica* 55.5, pp. 474–480.

Rantala, Anne, David P Fewer, Michael Hisbergues, Leo Rouhiainen, Jaana Vaitomaa, Thomas Börner, and Kaarina Sivonen (2004). “Phylogenetic evidence for the early evolution of microcystin synthesis.” In: *Proceedings of the National Academy of Sciences of the United States of America* 101.2, pp. 568–73.

Rinehart, Kenneth L., Michio Namikoshi, and Byoung W. Choi (1994). “Structure and biosynthesis of toxins from blue-green algae (cyanobacteria)”. In: *Journal of Applied Phycology* 6.2, pp. 159–176.

Rocap, Gabrielle, Frank W Larimer, Jane Lamerdin, Stephanie Malfatti, Patrick Chain, Nathan a Ahlgren, Andrae Arellano, Maureen Coleman, Loren Hauser, Wolfgang R Hess, Zackary I Johnson, Miriam Land, Debbie Lindell, Anton F Post, Warren Regala, Manesh Shah, Stephanie L Shaw, Claudia Steglich, Matthew B Sullivan, Claire S Ting, Andrew Tolonen, Eric a Webb, Erik R Zinser, and Sallie W Chisholm (2003). “Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation”. In: *Nature* 424.6952, pp. 1042–1047.

Rocha, Eduardo P C (2006). “Inference and analysis of the relative stability of bacterial chromosomes”. In: *Molecular Biology and Evolution* 23.3, pp. 513–522.

Rounge, Trine B, Thomas Rohrlack, Alexander J Nederbragt, Tom Kristensen, and Kjetill S Jakobsen (2009). “A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a Planktothrix rubescens strain.” In: *BMC Genomics* 10, p. 396.

Sambrook, Joseph and David William Russell (2001). *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, p. 766.

Sattath, Shmuel, Eyal Elyashiv, Oren Kolodny, Yosef Rinott, and Guy Sella (2011). “Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans”. In: *PLoS Genetics* 7.2, pp. 24–29.

Schatz, Daniella, Yael Keren, Assaf Vardi, Assaf Sukenik, Shmuel Carmeli, Thomas Börner, Elke Dittmann, and Aaron Kaplan (2007). “Towards clarification of the biological role of microcystins, a family of cyanobacterial toxins”. In: *Environmental Microbiology* 9.4, pp. 965–970.

Schmidt, K, S Mwaigwisya, L C Crossman, M Doumith, D Munroe, C Pires, A M Khan, N Woodford, N J Saunders, J Wain, J. O’Grady, and D M Livermore (2017a). “Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing”. In: *Journal of Antimicrobial Chemotherapy* 72.1, pp. 104–114.

Schmidt, Maximilian H.-W., Alxander Vogel, Alisandra Denton, Benjamin Istace, Alexandra Wormit, Henri van de Geest, Marie E. Bolger, Saleh Alseekh, Janina Mass, Christian Pfaff, Ulrich Schurr, Roger Chetelat, Florian Maumus, Jean-Marc Aury, Alisdair R. Fernie, Dani Zamir, Anthony M. Bolger, and Bjoern Usadel (2017b). "Reconstructing The Gigabase Plant Genome Of Solanum pennellii Using Nanopore Sequencing". In: *BioRxiv*, pp. 1–23.

Sedlazeck, Fritz J, Philipp Rescheneder, Moritz Smolka, Han Fang, and Maria Nattestad (2017). "Accurate detection of complex structural variations using single molecule sequencing". In: *bioRxiv*, pp. 1–24.

Shi, Tuo and Paul G Falkowski (2008). "Genome evolution in cyanobacteria: the stable core and the variable shell." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.7, pp. 2510–2515.

Silva, Diogo N, Sebastien Duplessis, Pedro Talhinhas, Helena Azinheira, Octávio S. Paulo, and Dora Batista (2015). "Genomic Patterns of Positive Selection at the Origin of Rust Fungi". In: *PLoS ONE* 10.12. Ed. by Marc Robinson-Rechavi, e0143959.

Sivonen, Kaarina (1996). "Cyanobacterial toxins and toxin production". In: *Phycol* 35.6, pp. 12–24.

Sović, Ivan, Mile Šikić, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen, and Niranjan Nagarajan (2016). "Fast and sensitive mapping of nanopore sequencing reads with GraphMap". In: *Nature Communications* 7, p. 11307. arXiv: `9809069v1 [arXiv:gr-qc]`.

Stucken, Karina, Uwe John, Allan Cembella, Alejandro A. Murillo, Katia Soto-Liebe, Juan J. Fuentes-Valdés, Maik Friedel, Alvaro M. Plominsky, Mónica Vásquez, and Gernot Glöckner (2010). "The smallest known genomes of multicellular and toxic cyanobacteria: Comparison, minimal gene sets for linked traits and the evolutionary implications". In: *PLoS ONE* 5.2.

Tillett, Daniel, Elke Dittmann, Marcel Erhard, Hans Von Döhren, Thomas Börner, and Brett A. Neilan (2000). "Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: An integrated peptide-polyketide synthetase system". In: *Chemistry and Biology* 7.10, pp. 753–764.

Tomitani, Akiko, Andrew H Knoll, Colleen M Cavanaugh, and Terufumi Ohno (2006). "The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.14, pp. 5442–5447.

Tooming-Klunderud, Ave, Hanne Sogge, Trine Ballestad Rounge, Alexander J. Nederbragt, Karin Lagesen, Gernot Glöckner, Paul K. Hayes, Thomas Rohrlack, and Kjetill S. Jakobsen (2013). "From green to red: Horizontal gene transfer of the phycoerythrin gene cluster between Planktothrix strains". In: *Applied and Environmental Microbiology* 79.21, pp. 6803–6812.

Valério, Elisabete, Lélia Chambel, Sérgio Paulino, Natália Faria, Paulo Pereira, and Rogério Tenreiro (2009). "Molecular identification, typing and traceability of cyanobacteria from freshwater reservoirs". In: *Microbiology* 155.2, pp. 642–656.

Wright, G D (2007). "The antibiotic resistome: the nexus of chemical and genetic diversity". In: *Nat.Rev.Microbiol.* 5.1740-1534 (Electronic), pp. 175–186.

# 6 Supplementary Materials

Table S1: Quality and quantity assessment of gDNA stock samples

| | Qubit | Nanodrop | | |
|---|---|---|---|---|
| Strain | Concentration(ng/$\mu$L) | Concentration(ng/$\mu$L) | 260/280 | 260/230 |
| LMECYA7 | 106 | 1607.00 | 1.91 | 1.68 |
| LMECYA167 | 73.2 | 192.5 | 1.84 | 2.03 |
| LMECYA269 | 98.4 | 214.9 | 1.81 | 1.56 |
| LMECYA280 | 53.2 | 236.9 | 2.08 | 2.33 |
| LEGE06226 | 49.4 | 99.4 | 1.93 | 1.98 |
| LEGE06233 | 24.0 | 130.1 | 1.96 | 1.94 |

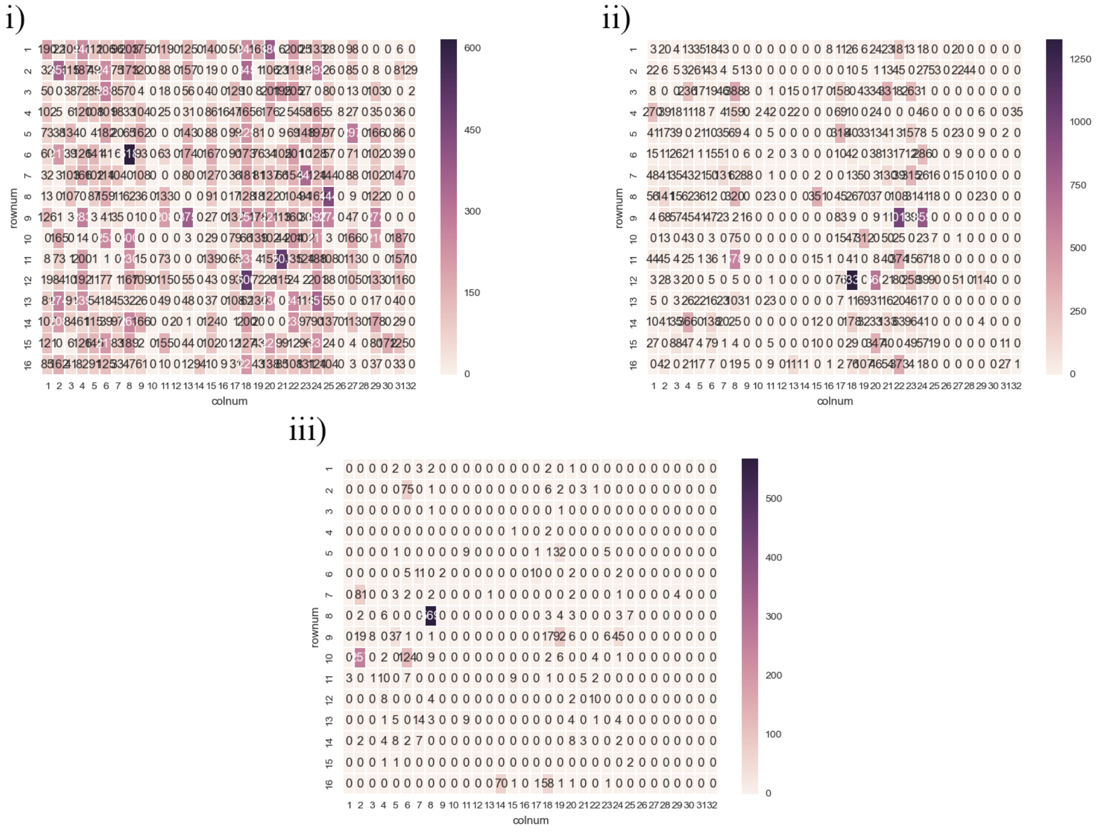

Figure S1: Pore occupancy during the first sequencing round. i) Lambda phage, ii) LMECYA167 and iii) LEGE06233.

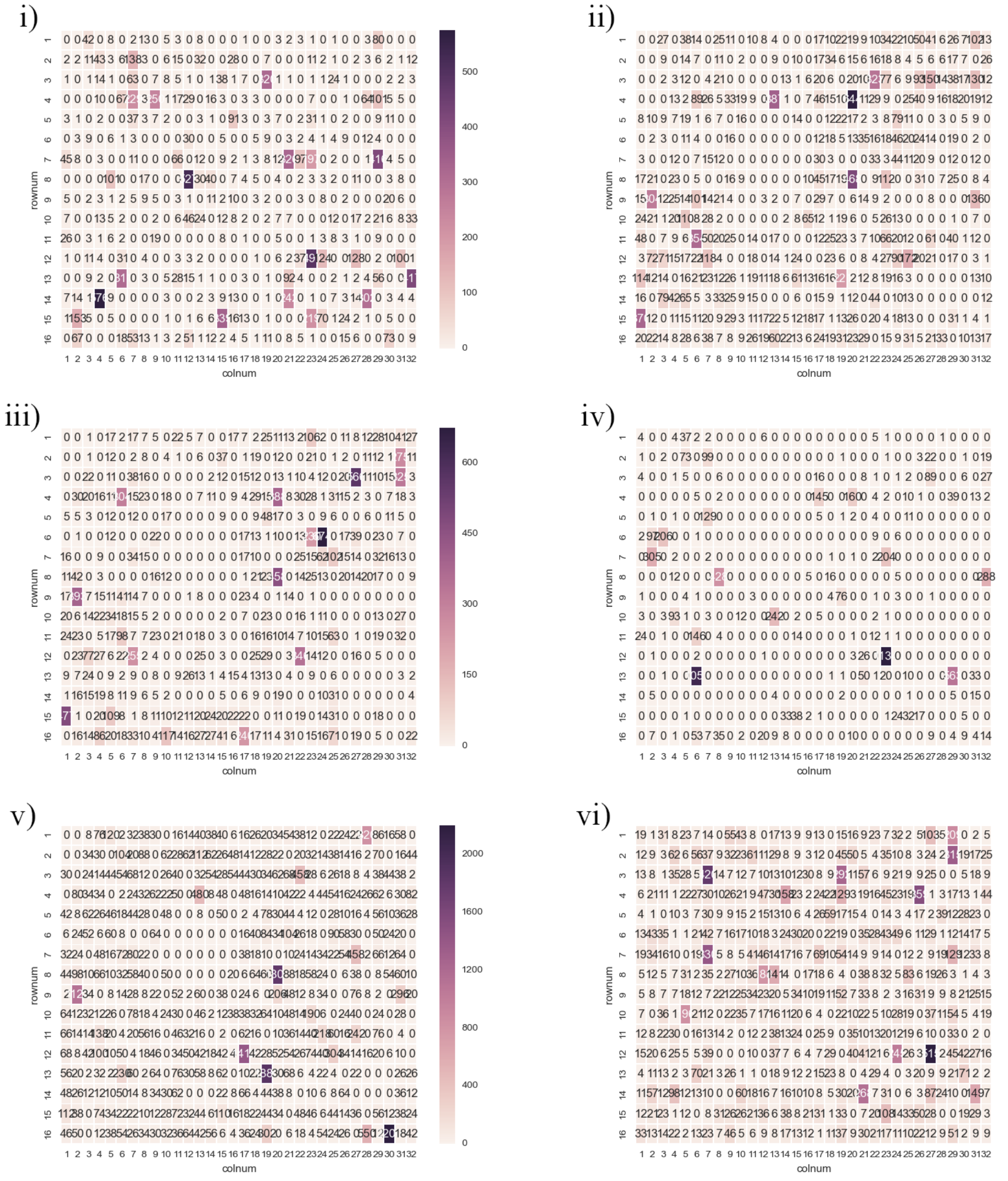

Figure S2: Pore occupancy during the second sequencing round. i) LMECYA7, ii) LMECYA167, iii) LMECYA269, iv) LMECYA280, v) LEGE06226 and vi) LEGE06233.

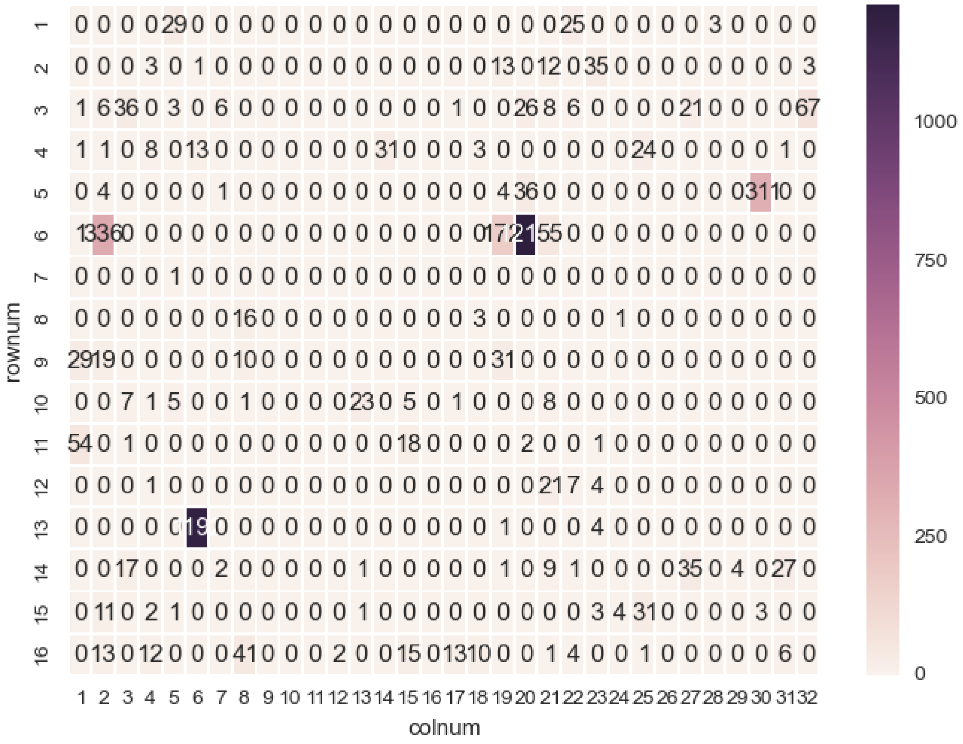

Figure S3: Pore occupancy during the LMECYA167 24h sequencing run.

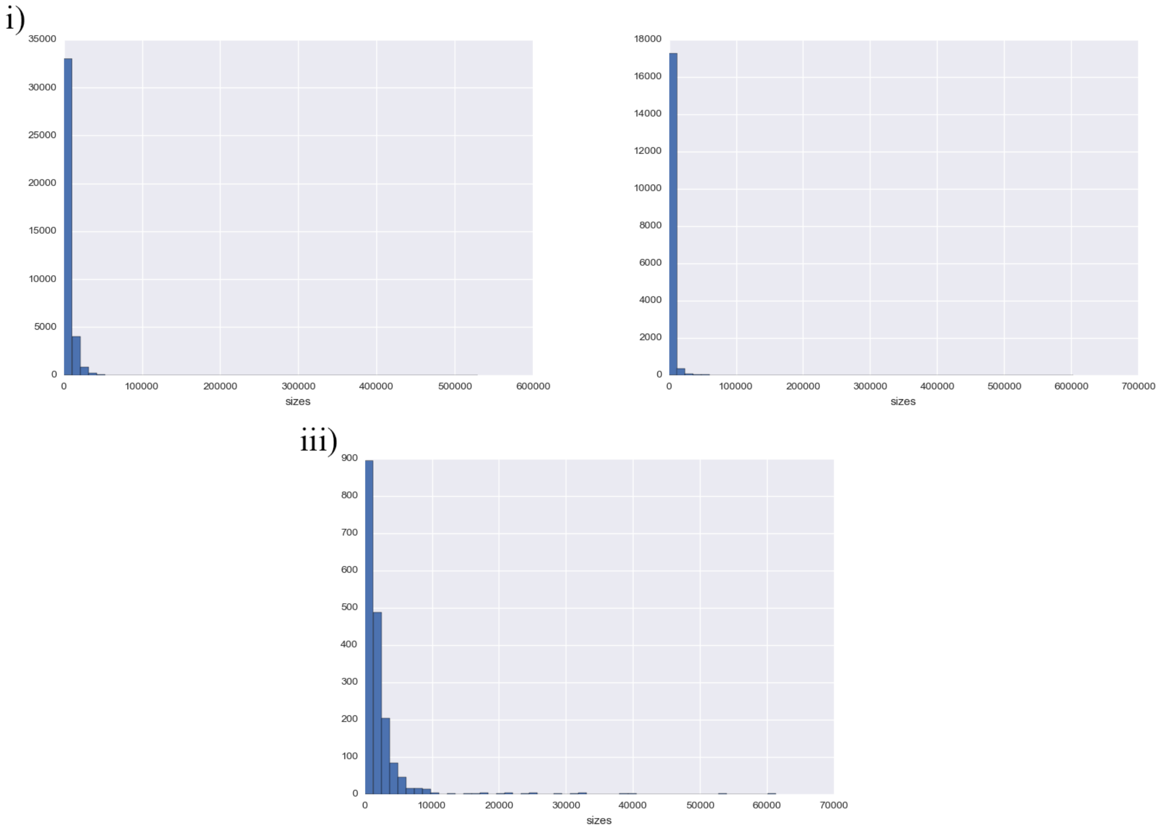

Figure S4: Read length during the first sequencing round. i) Lambda phage, ii) LMECYA167 and iii) LEGE06233.

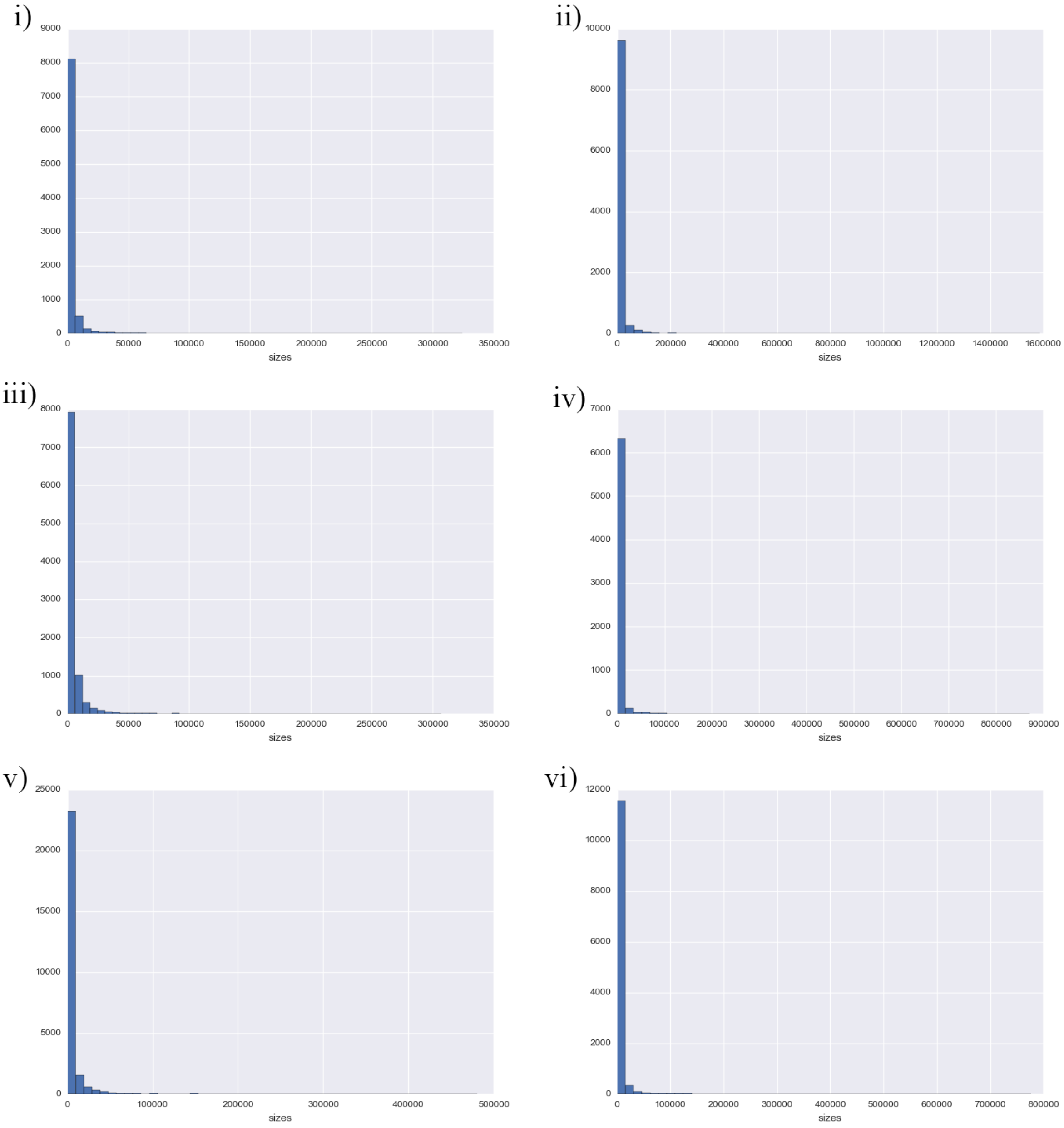

Figure S5: Read length during the second sequencing round. i) LMECYA7, ii) LMECYA167, iii) LMECYA269, iv) LMECYA280, v) LEGE06226 and vi) LEGE06233.

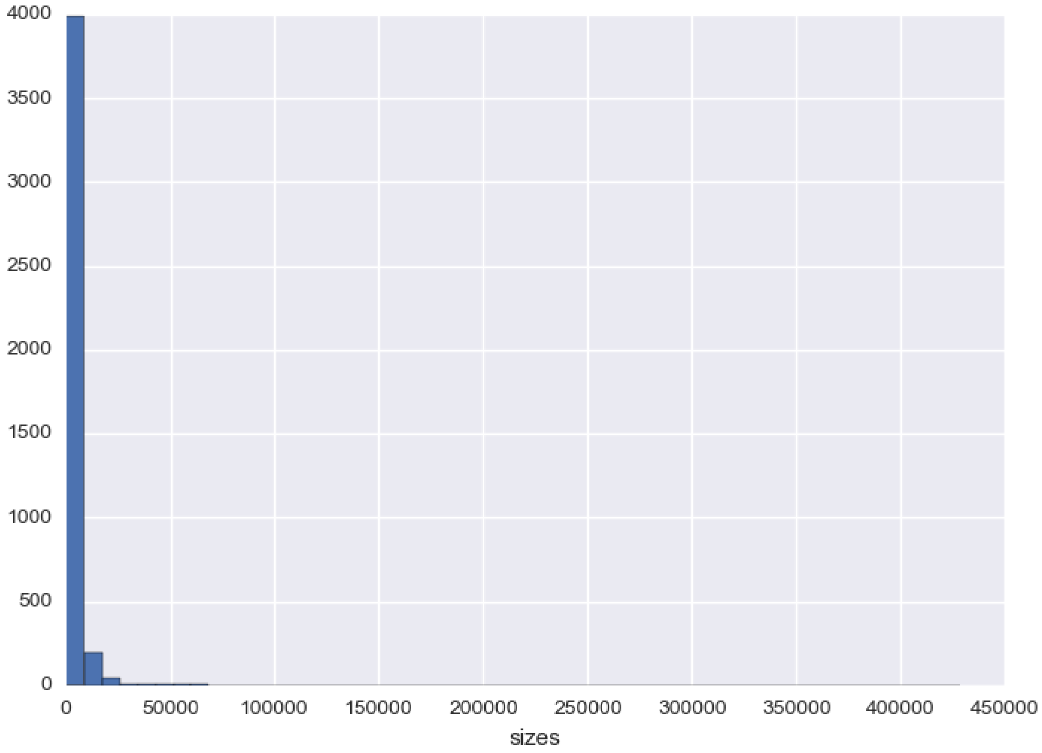

Figure S6: Read length during the LMECYA167 24h sequencing run.

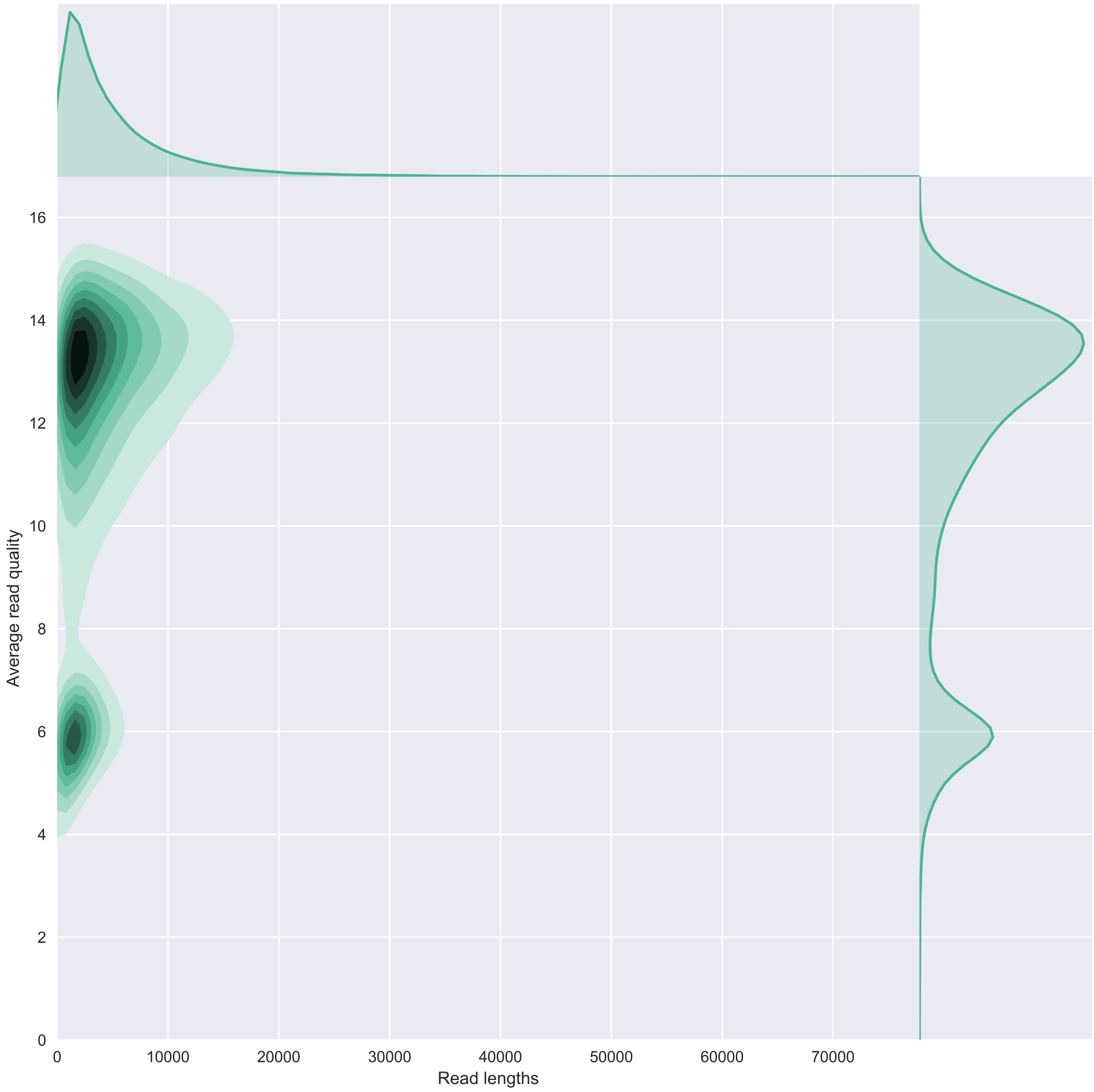

Figure S7: Lambda phage quality score versus read length plots. Albacore 1.2.1 basecalled reads.

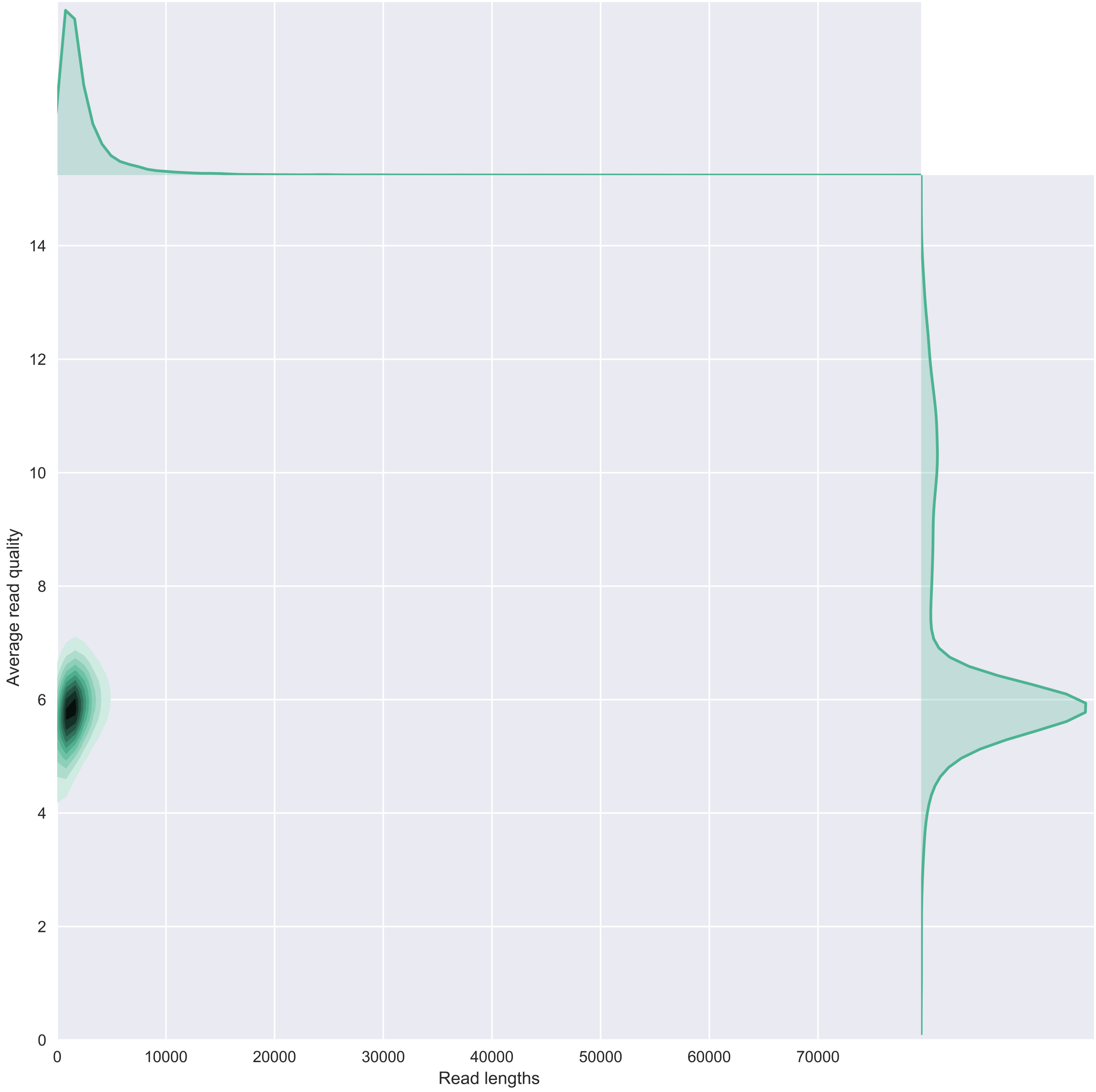

Figure S8: LMECYA167 quality score versus read length plots. Albacore 1.2.1 basecalled reads.

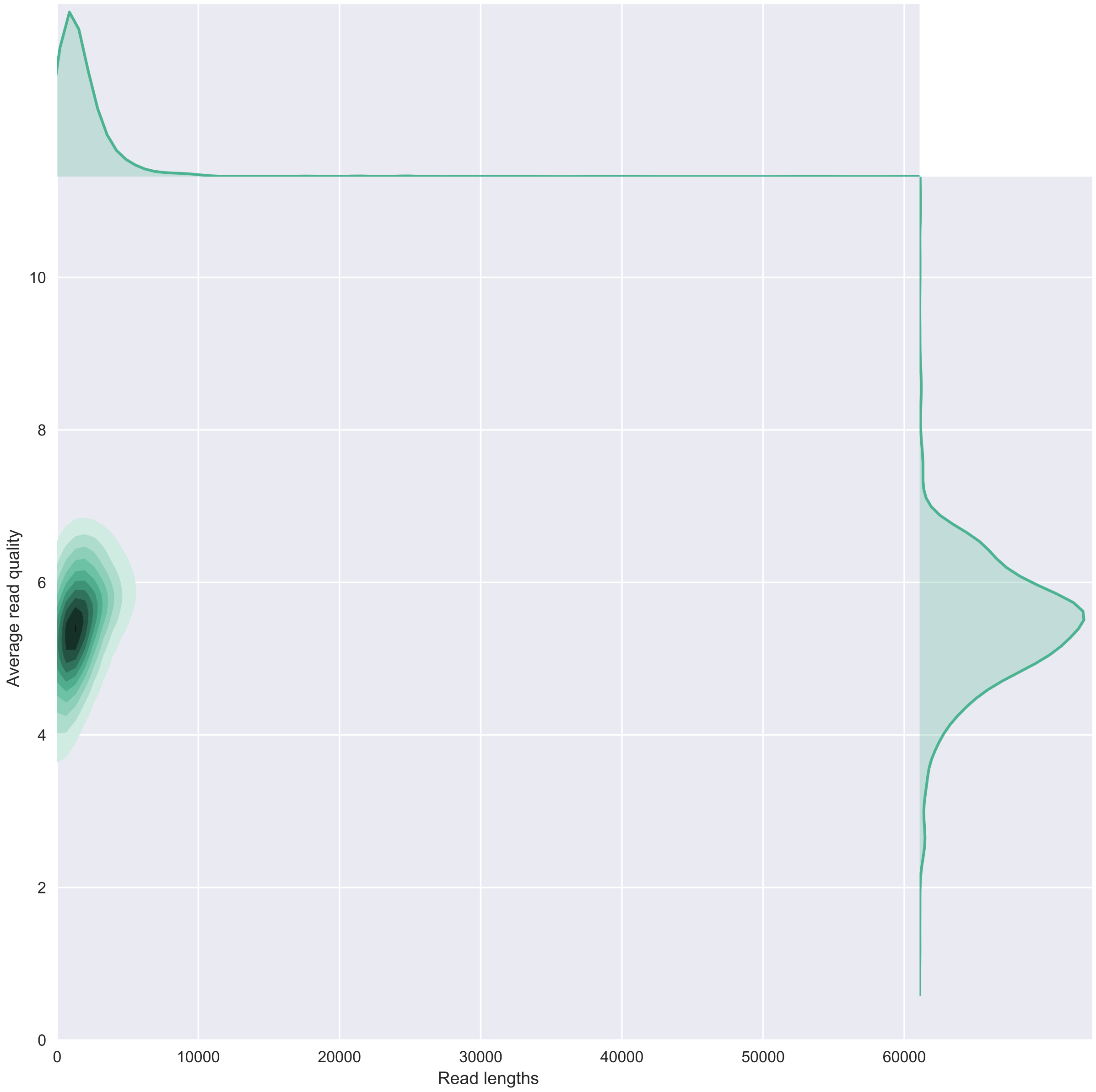

Figure S9: LEGE06233 quality score versus read length plots. Albacore 1.2.1 basecalled reads.

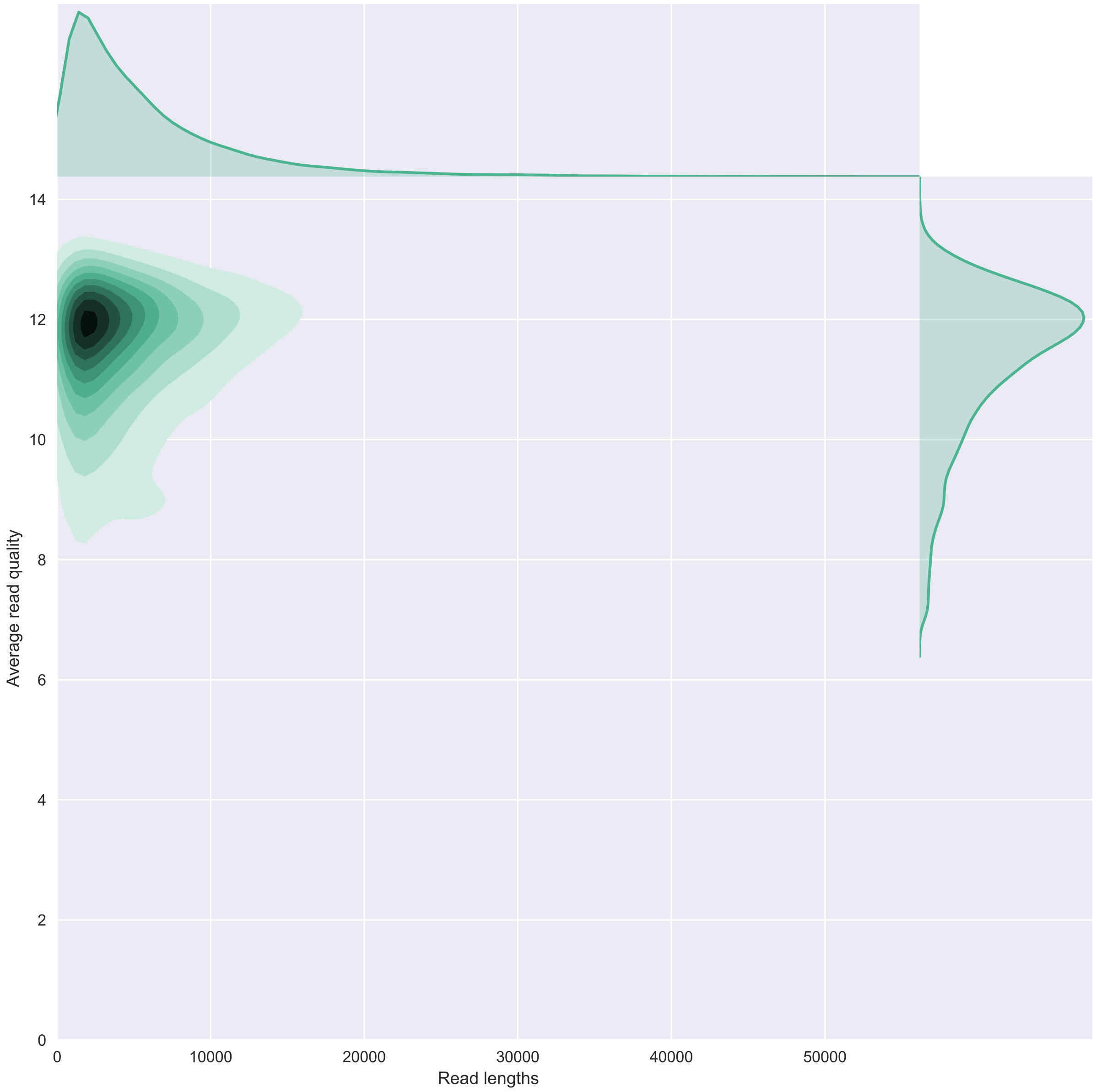

Figure S10: Lambda phage quality score versus read length plots. Albacore 1.2.1 basecalled reads.

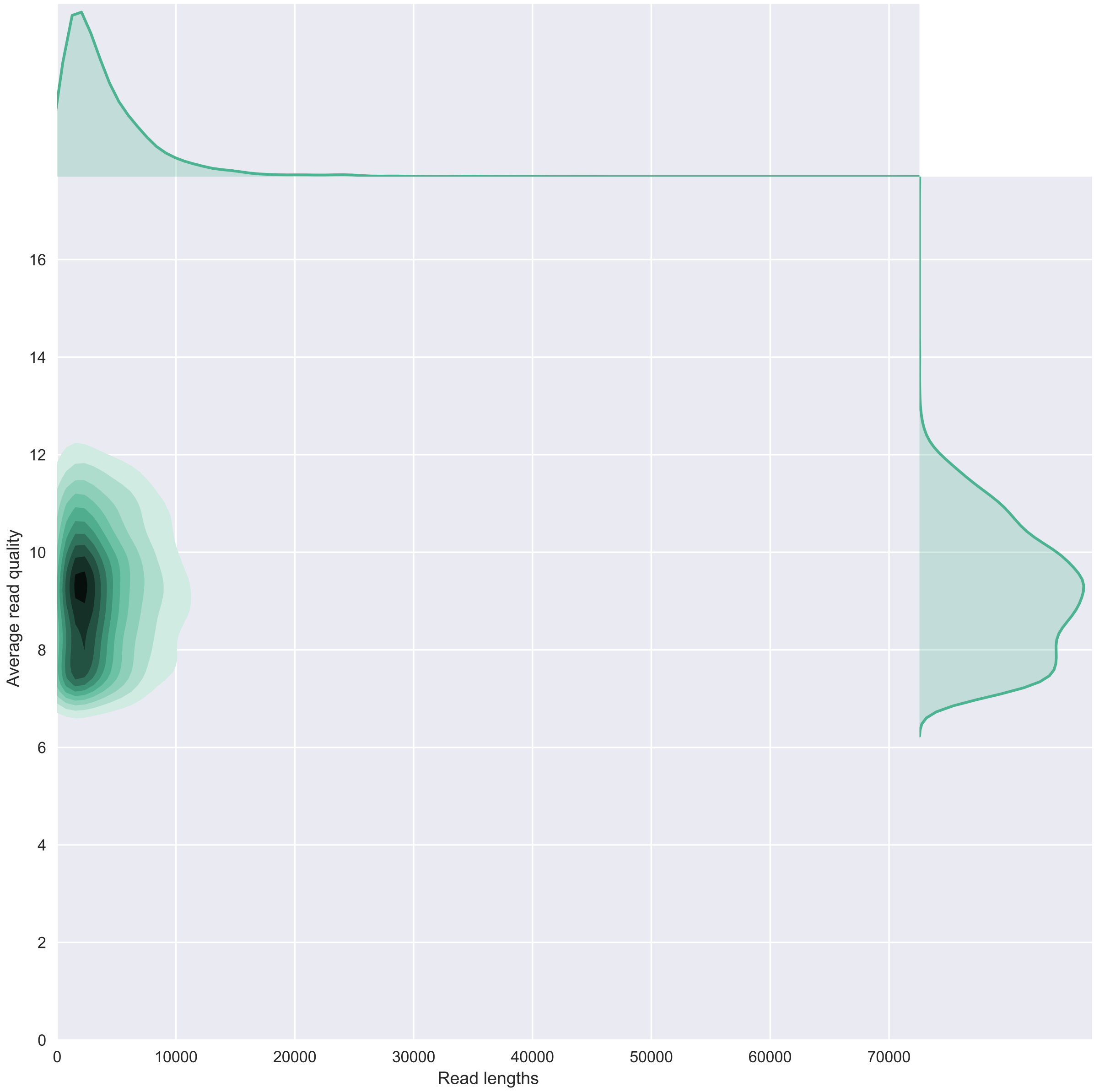

Figure S11: LMECYA167 quality score versus read length plots. Albacore 1.2.1 basecalled reads.

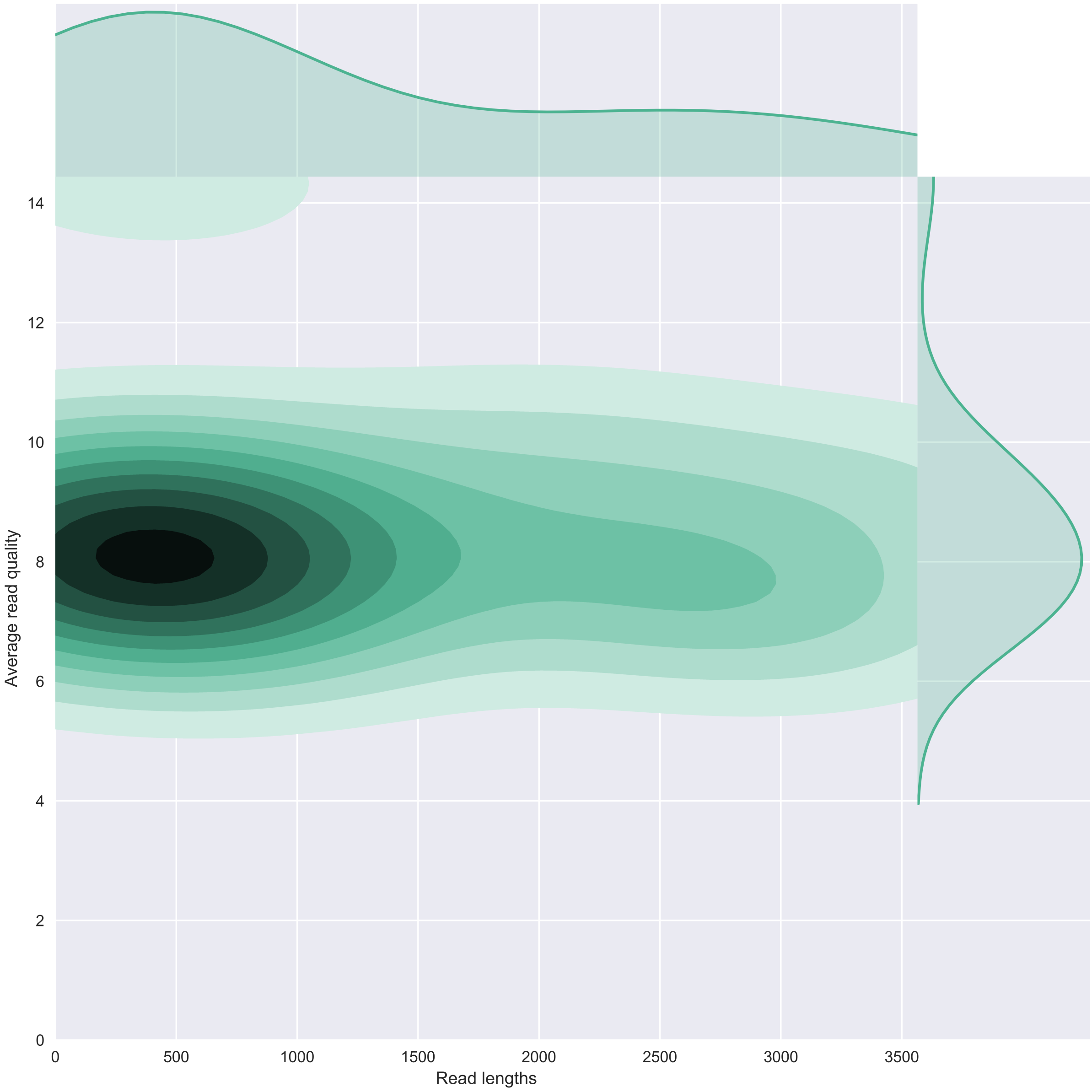

Figure S12: LEGE06233 quality score versus read length plots. Albacore 1.2.1 basecalled reads.

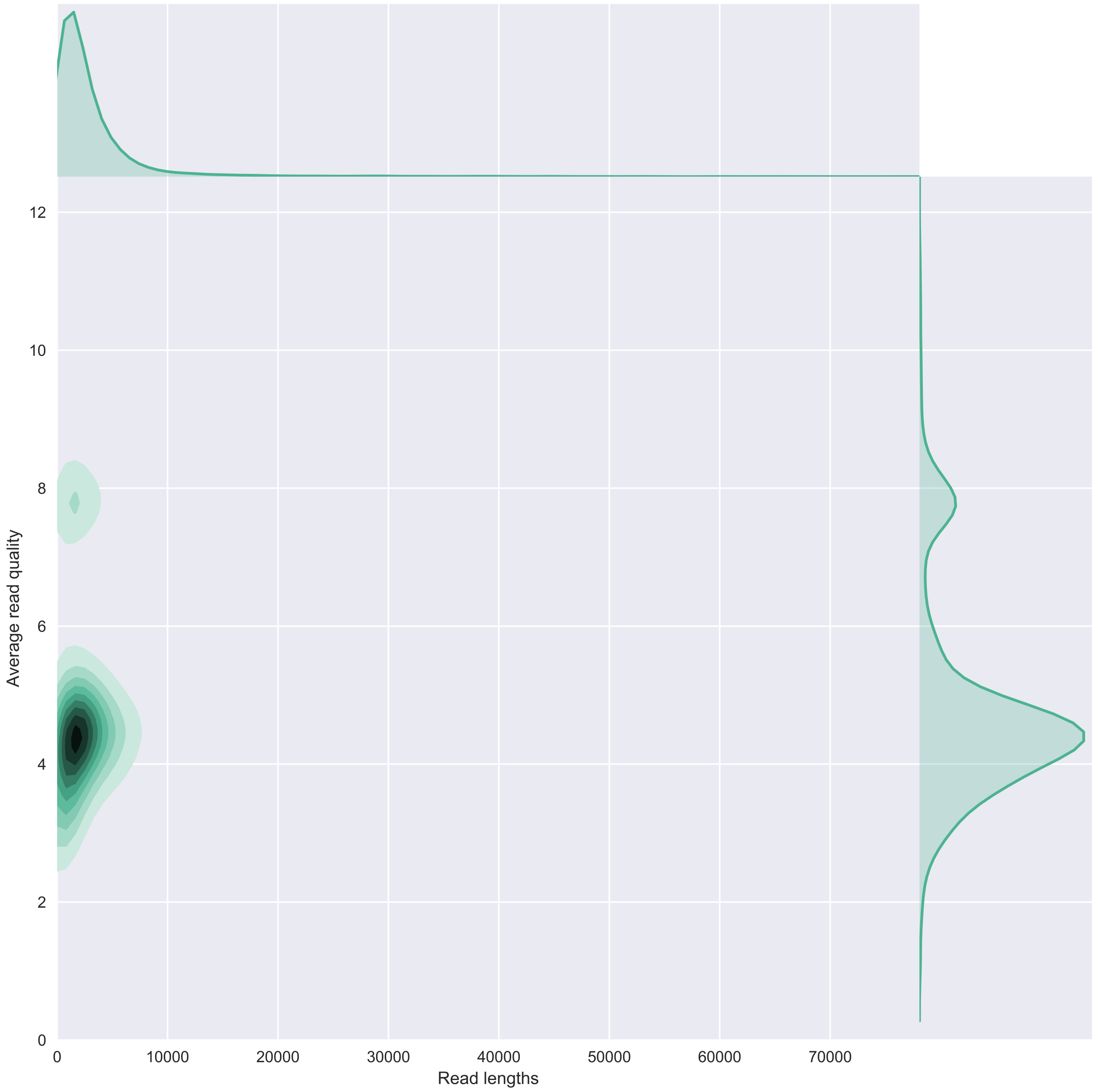

Figure S13: LMECYA7 quality score versus read length plots. Albacore 1.2.6 basecalled reads.

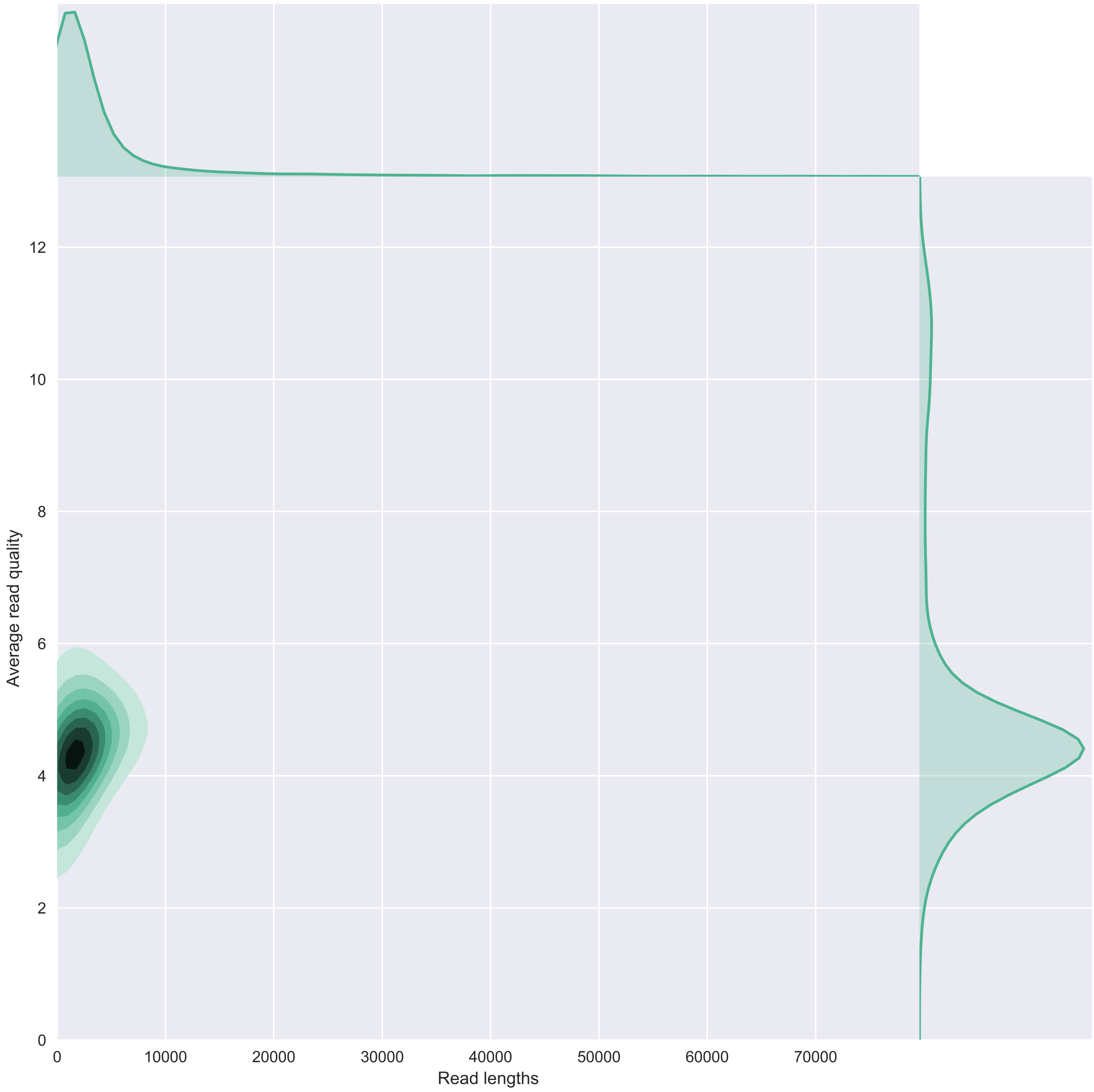

Figure S14: LMECYA167 quality score versus read length plots. Albacore 1.2.6 basecalled reads.

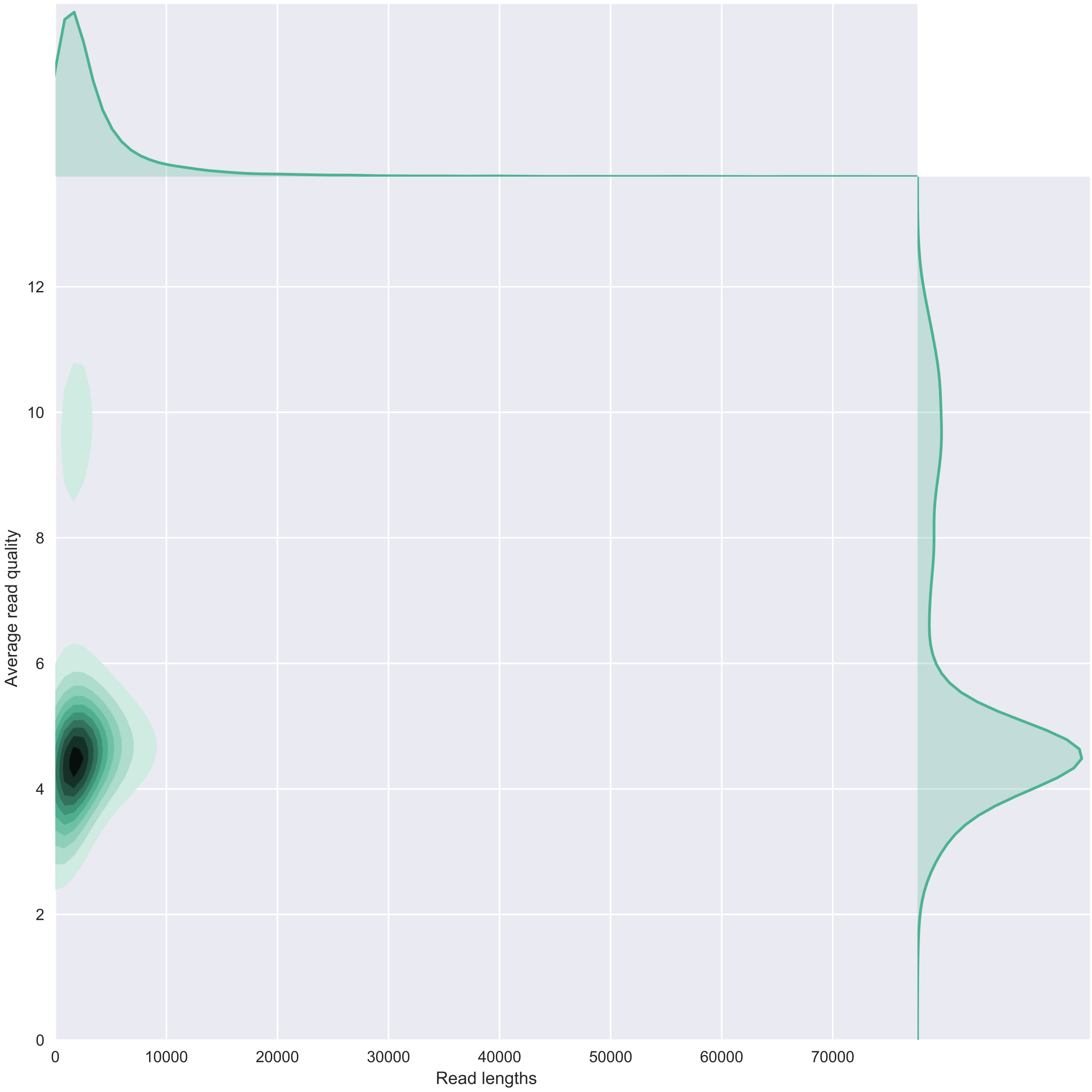

Figure S15: LMECYA269 quality score versus read length plots. Albacore 1.2.6 basecalled reads.

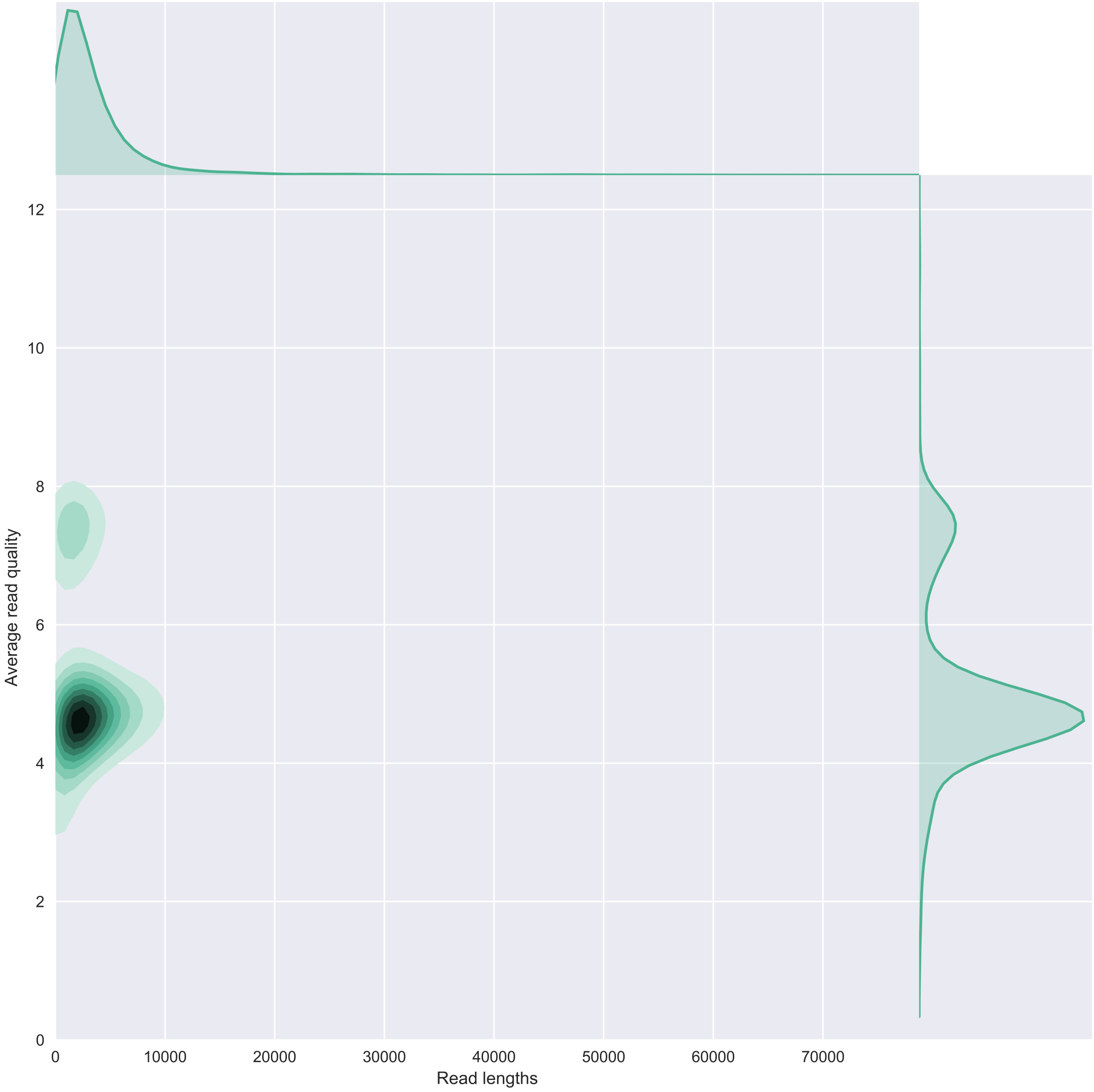

Figure S16: LMECYA280 quality score versus read length plots. Albacore 1.2.6 basecalled reads.

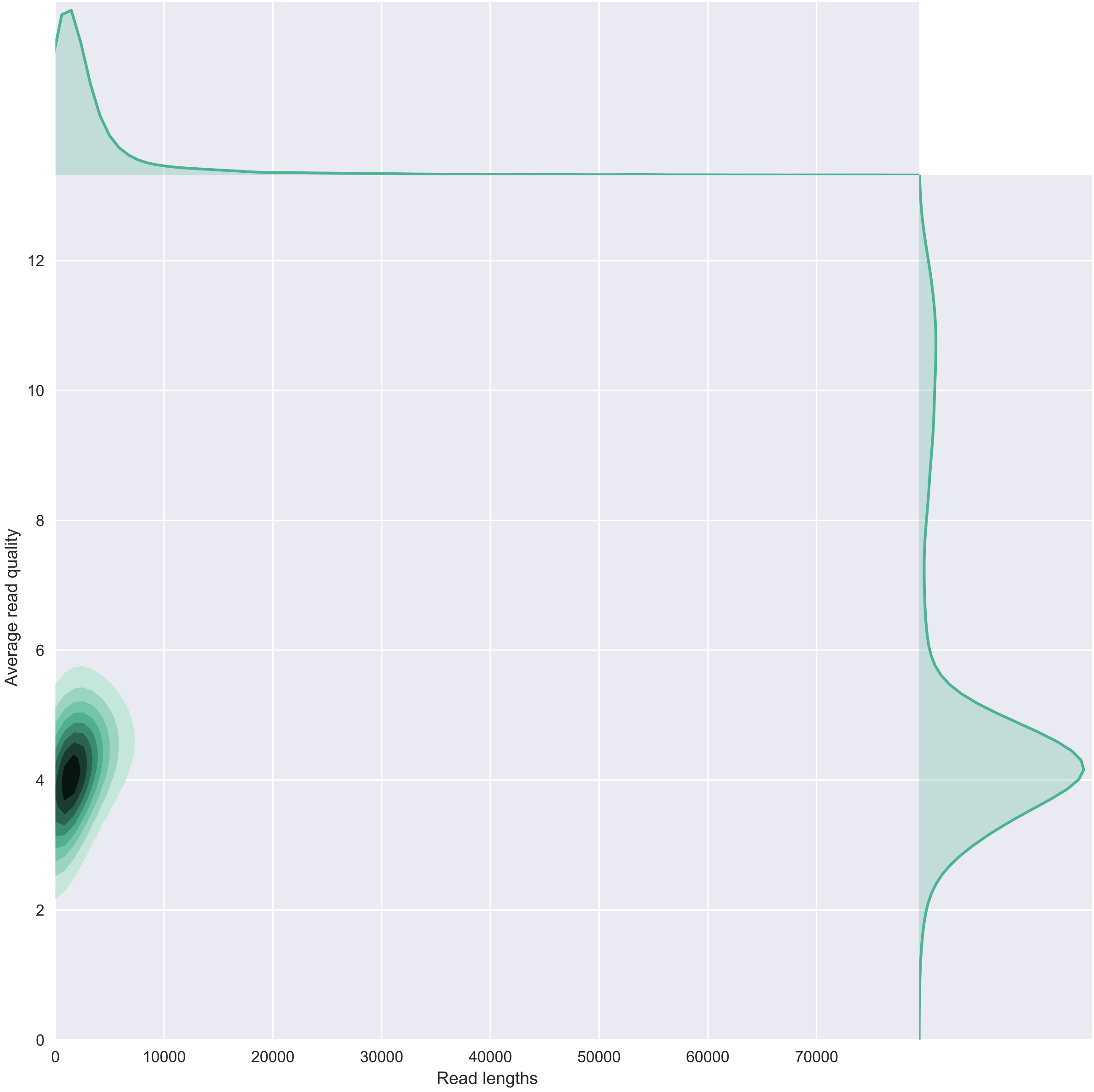

Figure S17: LEGE06226 quality score versus read length plots. Albacore 1.2.6 basecalled reads.

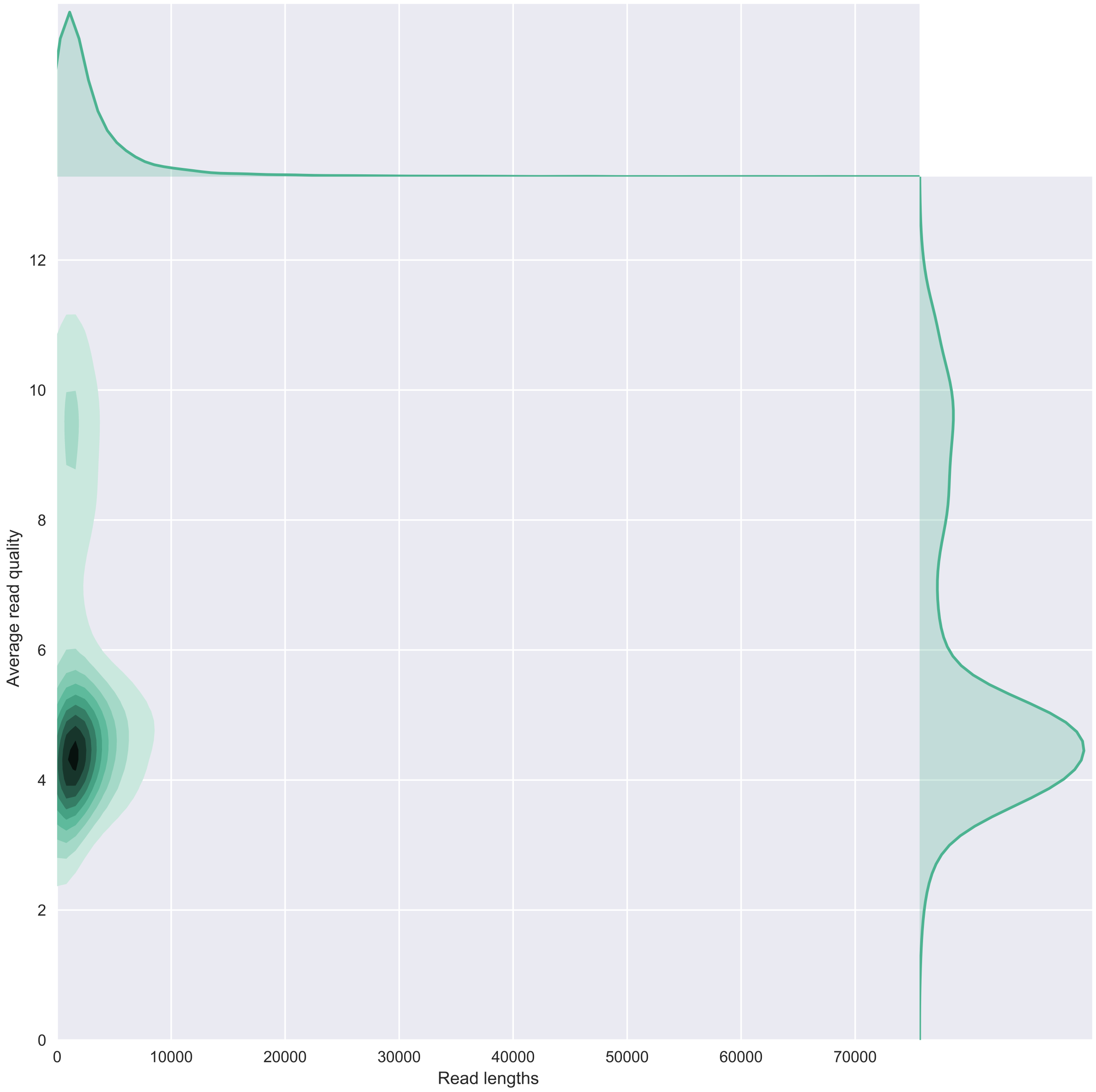

Figure S18: LEGE06233 quality score versus read length plots. Albacore 1.2.6 basecalled reads.

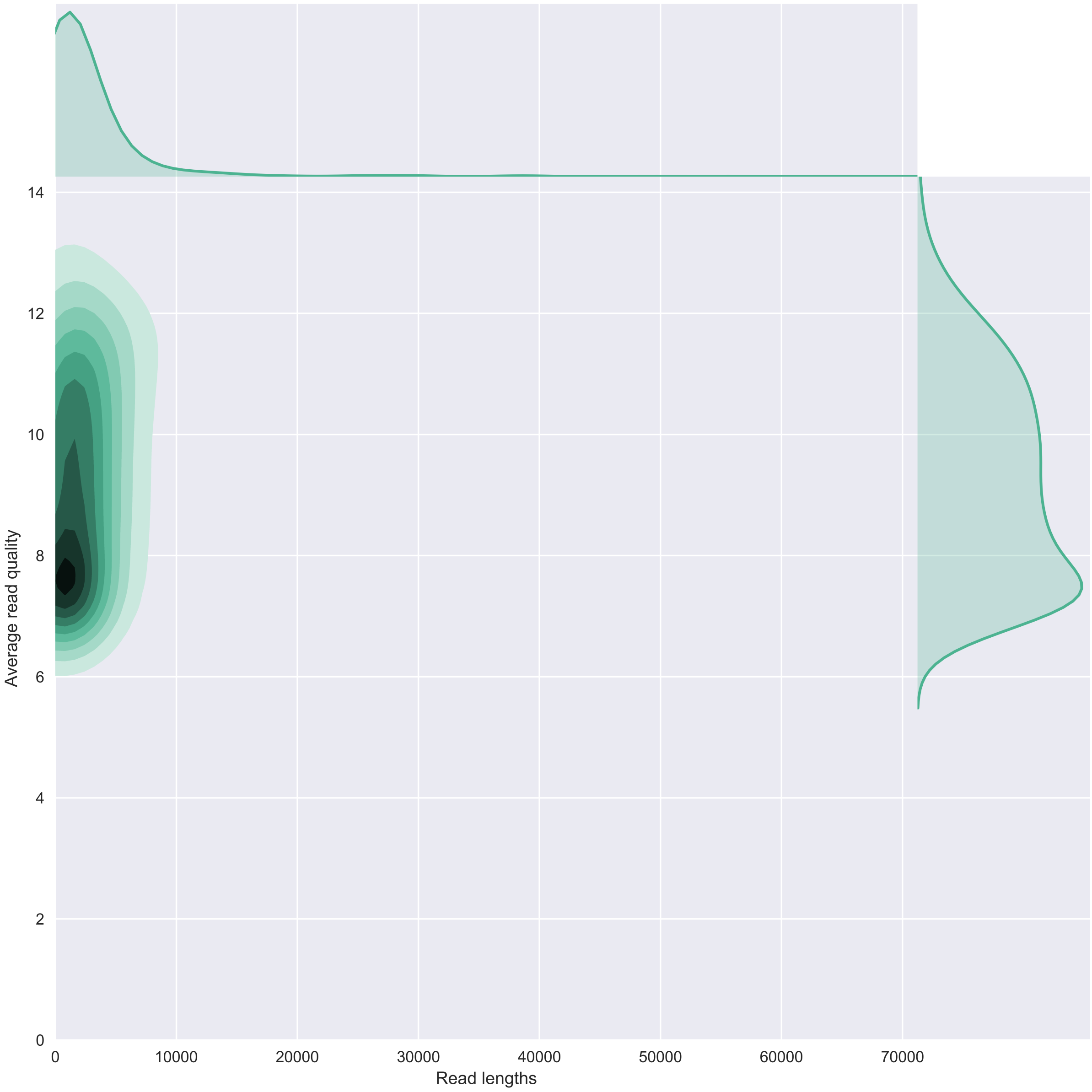

Figure S19: LMECYA7 quality score versus read length plots. Albacore 2.0.2 basecalled reads.

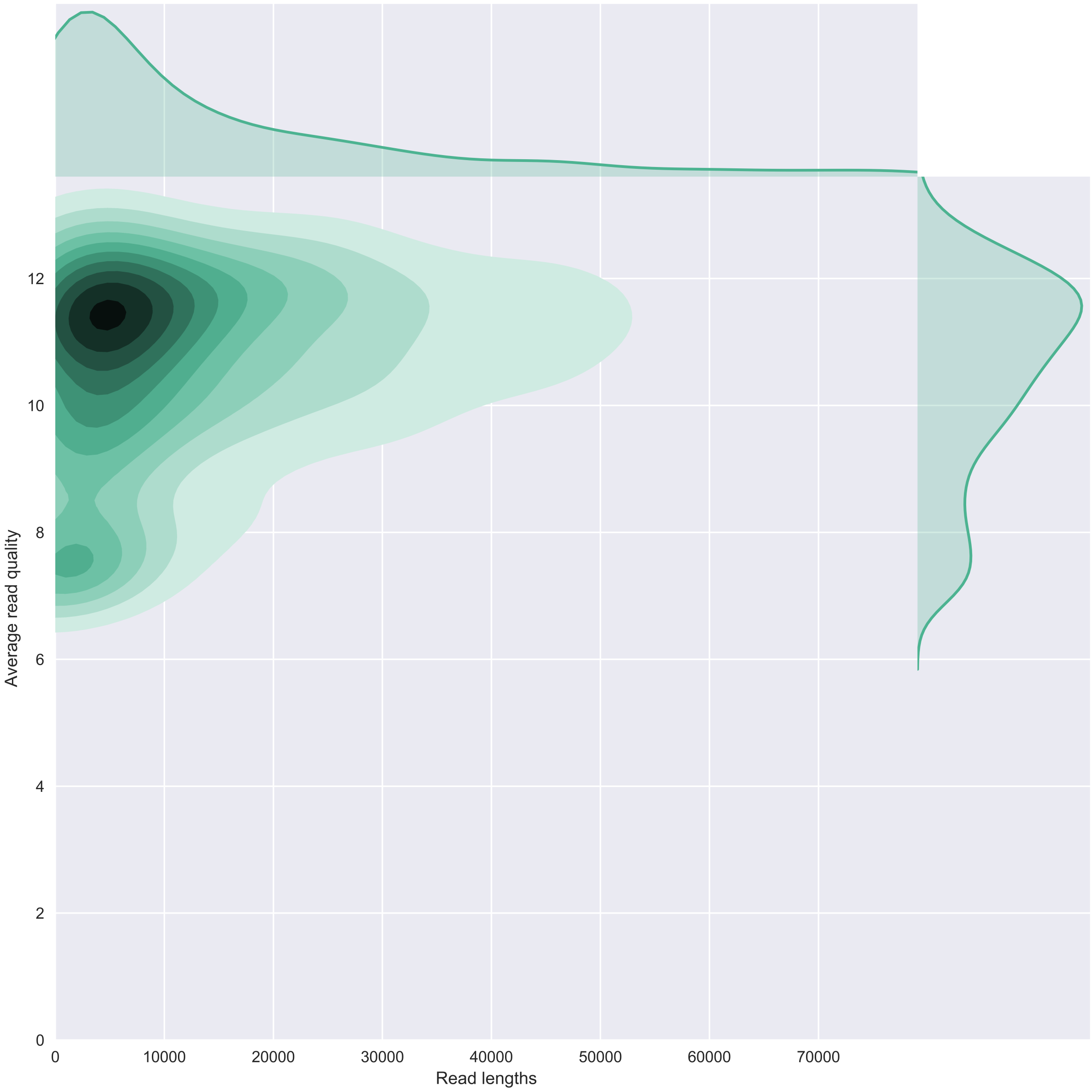

Figure S20: LMECYA167 quality score versus read length plots. Albacore 2.0.2 basecalled reads.

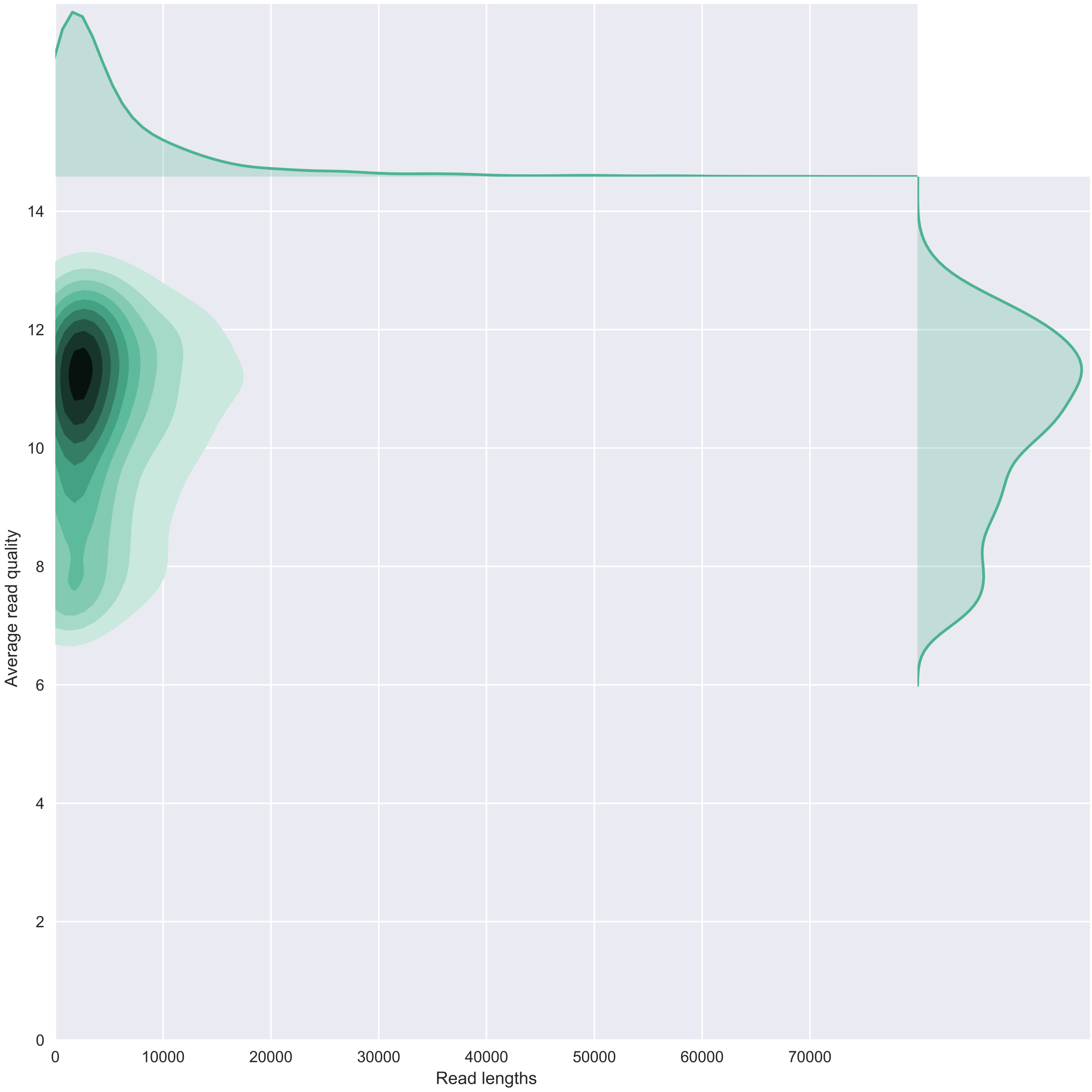

Figure S21: LMECYA269 quality score versus read length plots. Albacore 2.0.2 basecalled reads.

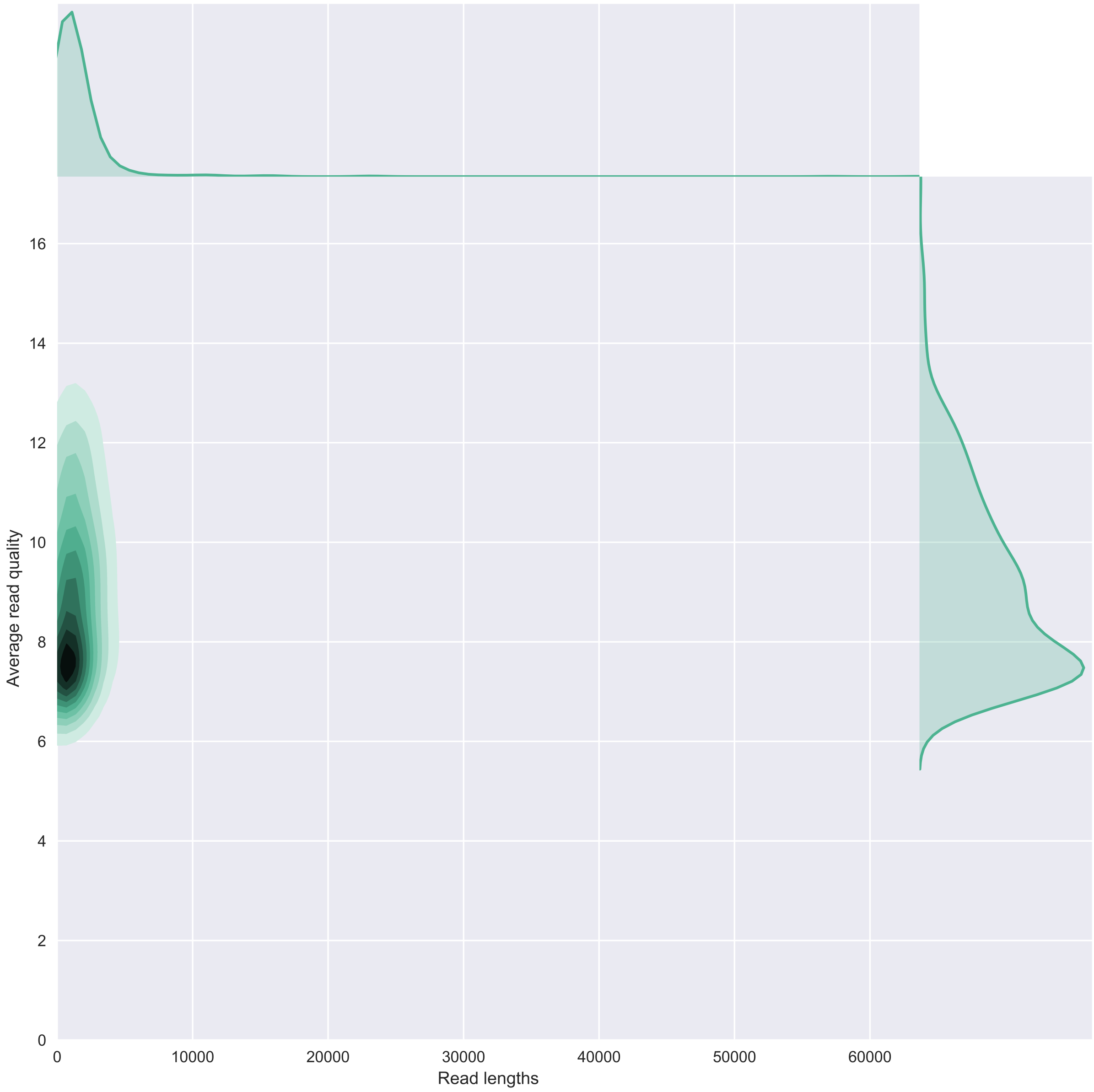

Figure S22: LMECYA280 quality score versus read length plots. Albacore 2.0.2 basecalled reads.

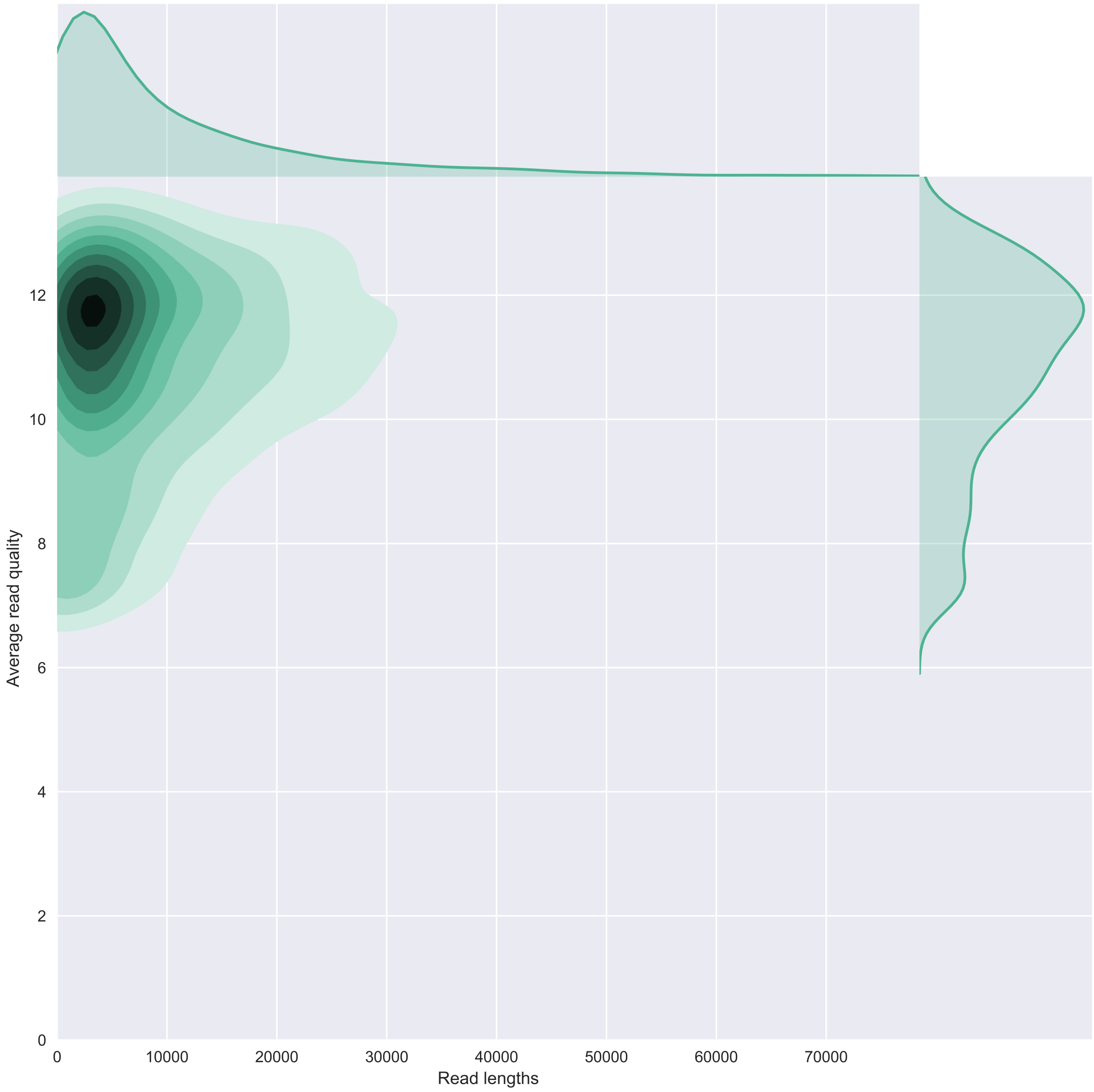

Figure S23: LEGE06226 quality score versus read length plots. Albacore 2.0.2 basecalled reads.

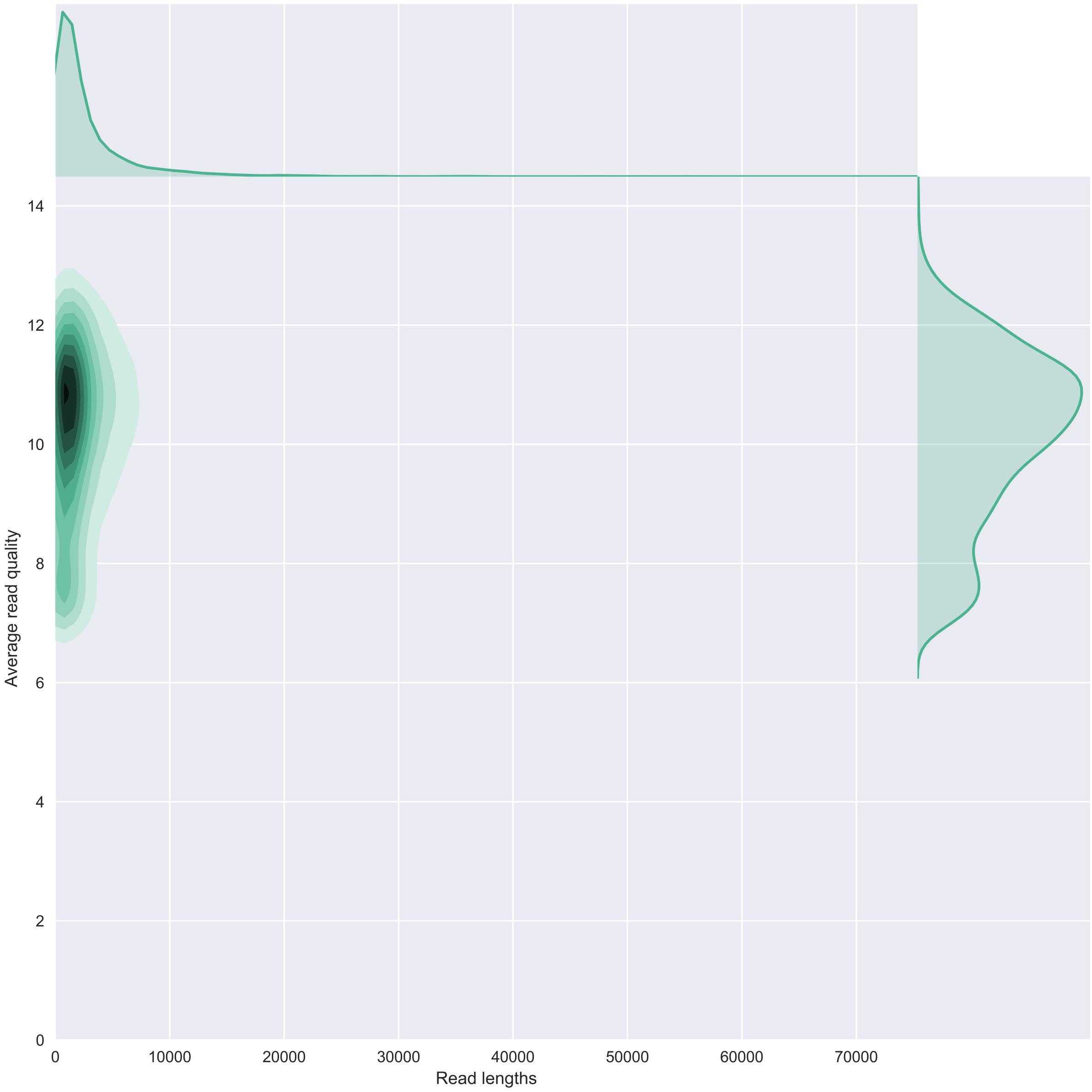

Figure S24: LEGE06233 quality score versus read length plots. Albacore 2.0.2 basecalled reads.

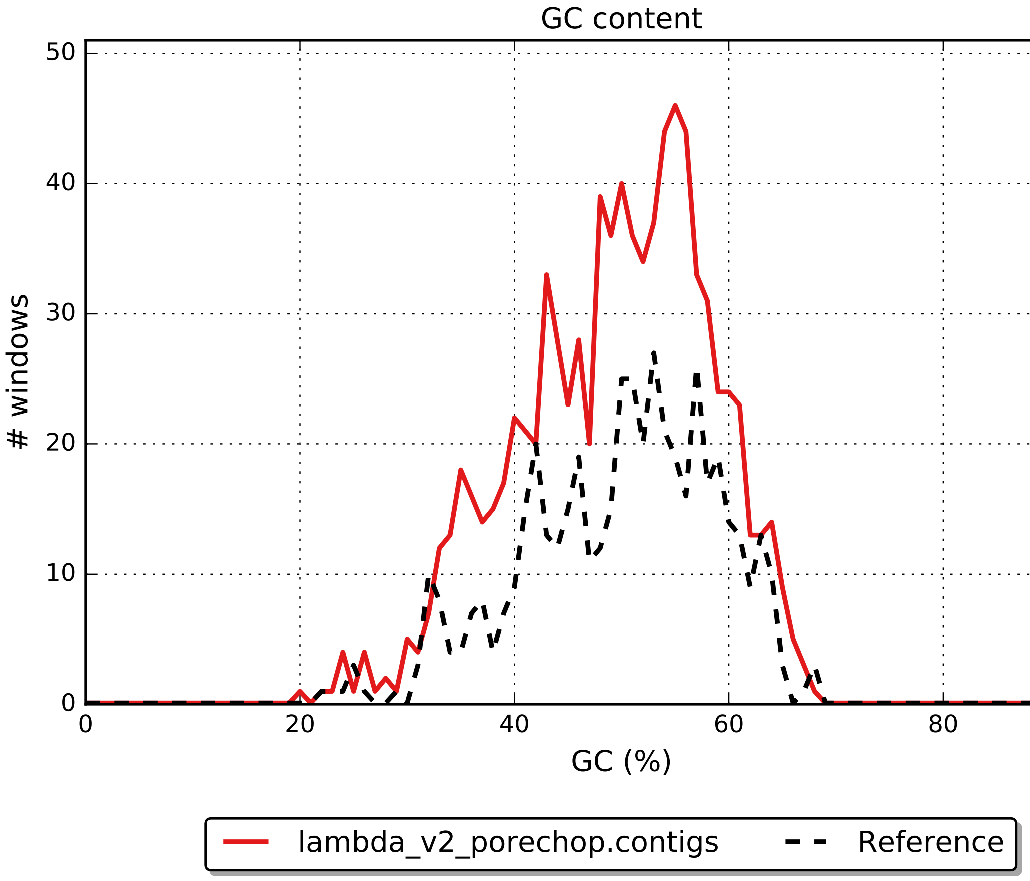

Figure S25: Lambda phage data assembly GC content pattern along the genome. Albacore 2.0.2 basecalled reads.

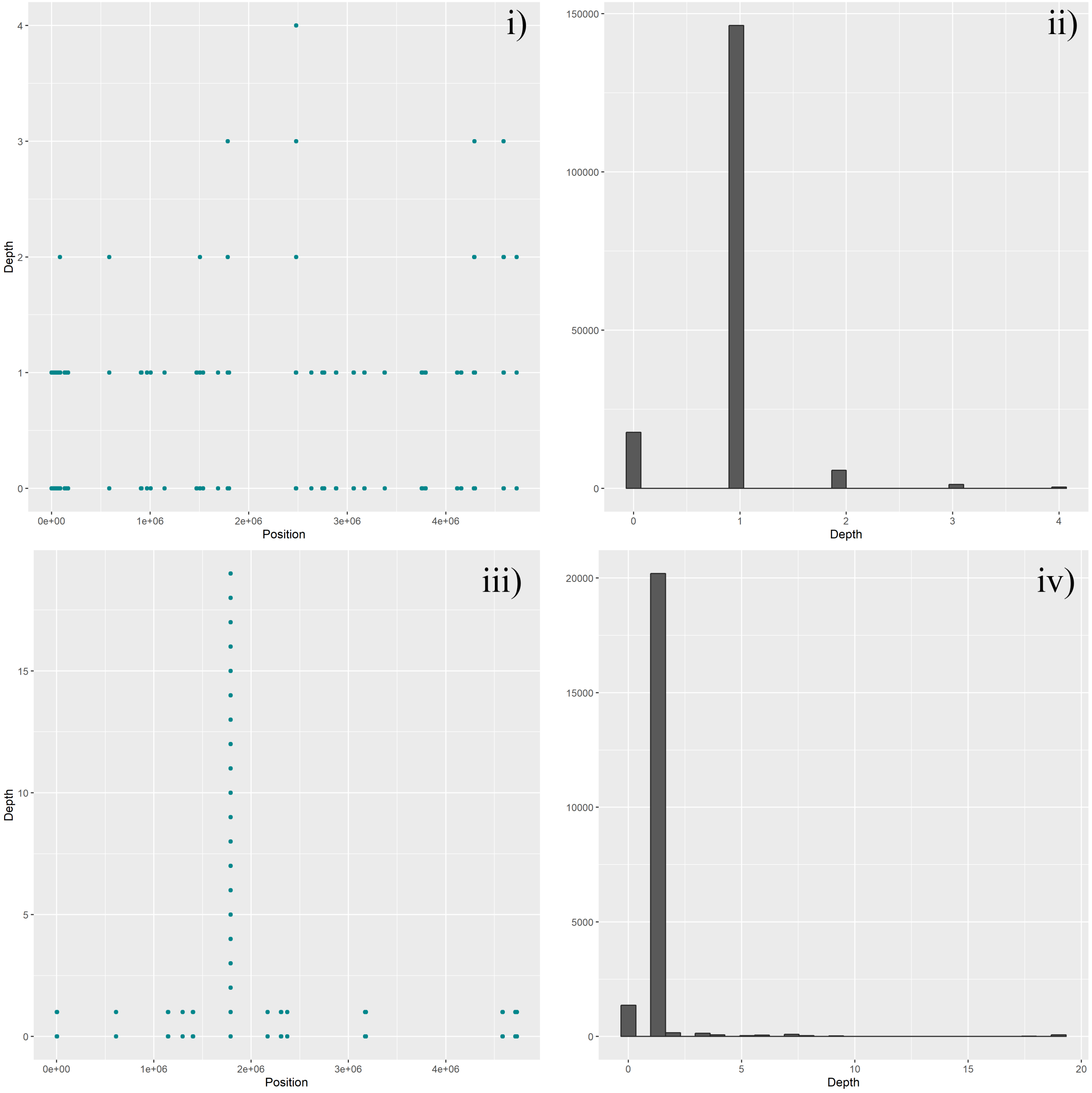

Figure S26: Sequencing depth along the genome (left hand-side plots) and overall distribution (right hand-side plots) of LMECYA269 (i and ii) and LMECYA280 (iii and iv).

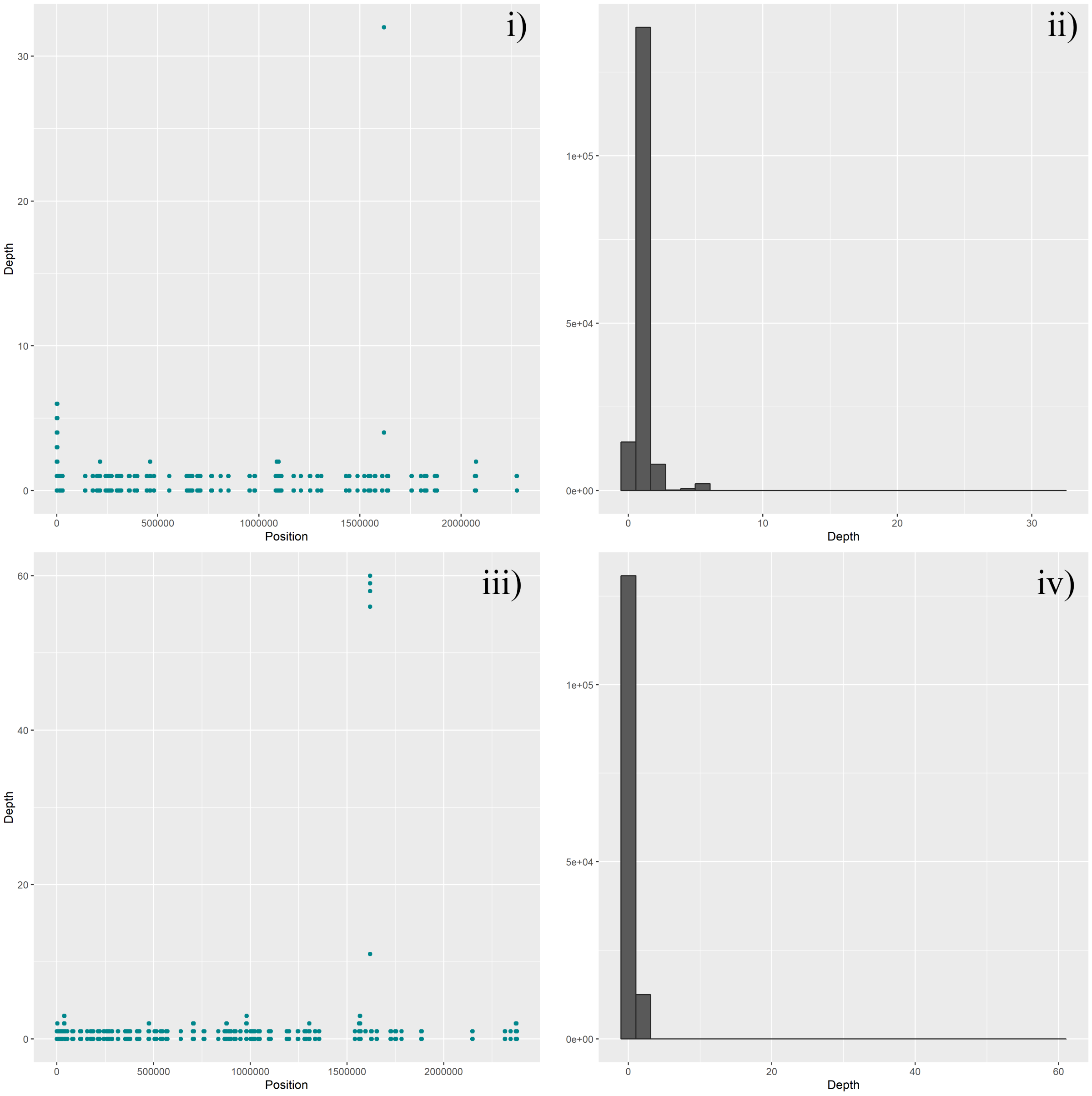

Figure S27: Sequencing depth along the genome (left hand-side plots) and overall distribution (right hand-side plots) of LEGE06226 (i and ii) and LEGE06233 (iii and iv).

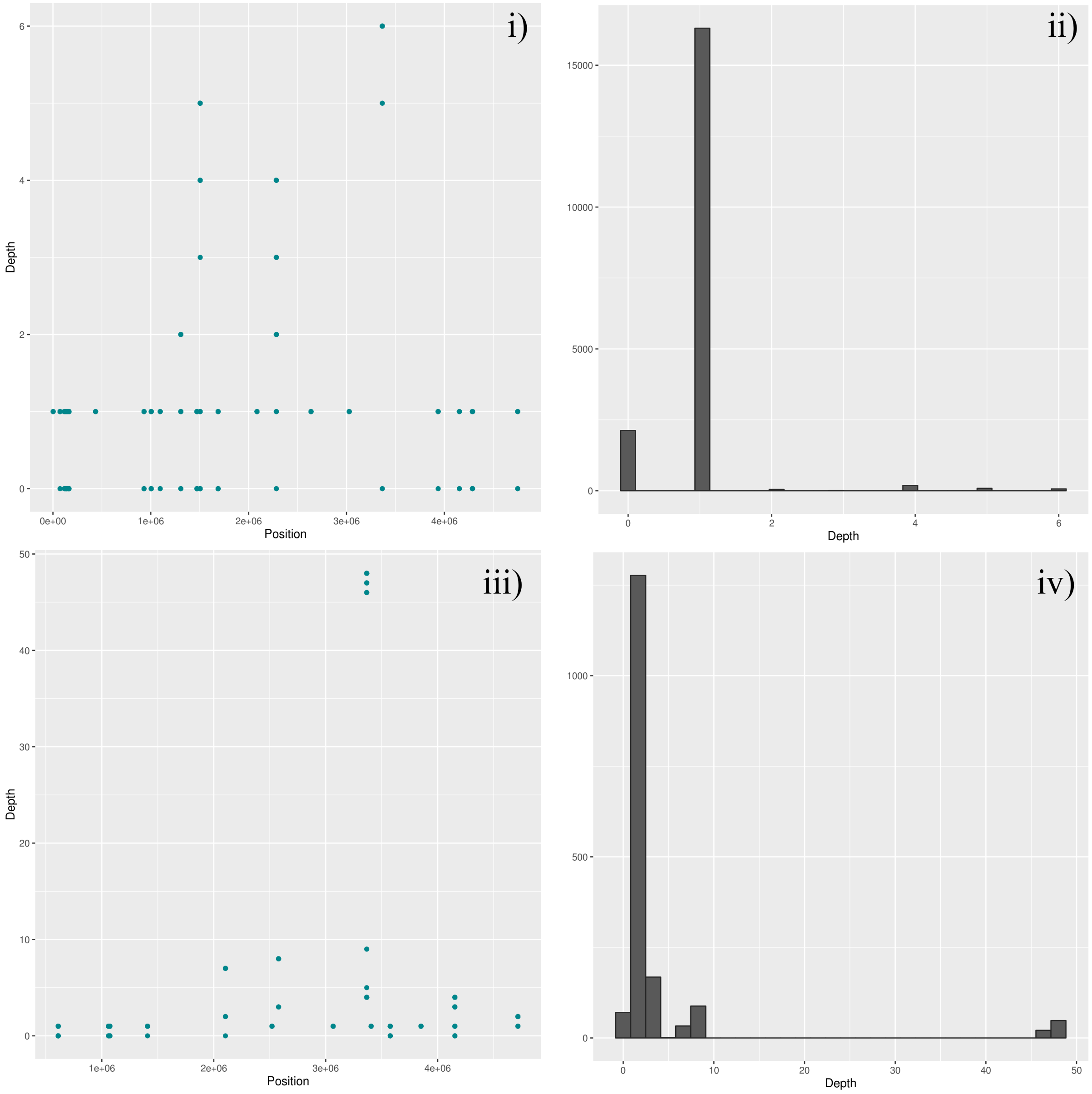

Figure S28: Sequencing depth along the genome (left hand-side plots) and overall distribution (right hand-side plots) of LMECYA269 (i and ii) and LMECYA280 (iii and iv). Alignment performed using Albacore 2.0.2 base-called reads.

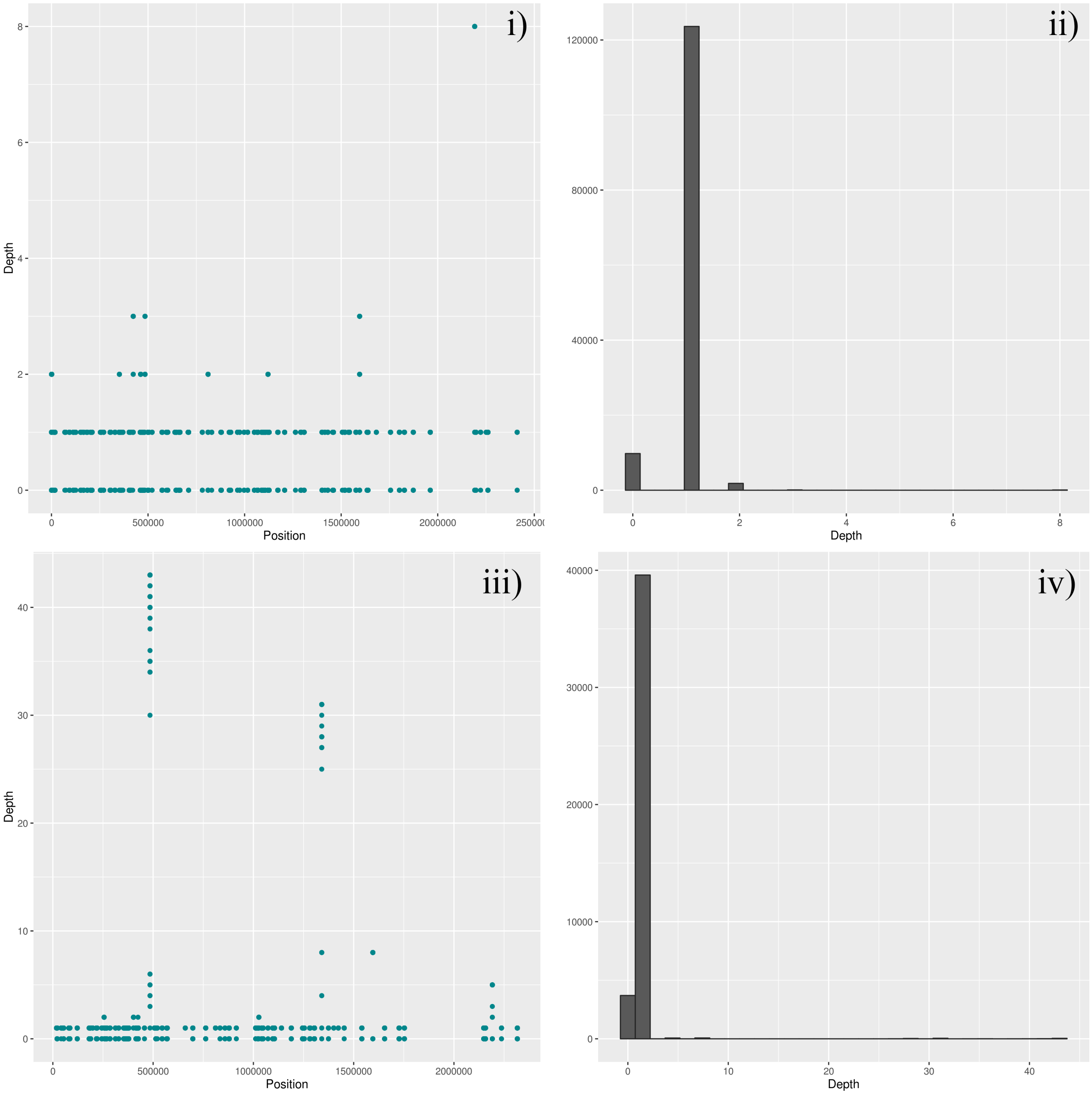

Figure S29: Sequencing depth along the genome (left hand-side plots) and overall distribution (right hand-side plots) of LEGE06226 (i and ii) and LEGE06233 (iii and iv). Alignment performed using Albacore 2.0.2 basecalled reads.

Table S2: LMECYA7 hybrid assembly alignment consensus RGI tool hits.

| **Start position (bp)** | **Stop position (bp)** | **Orientation** | **Best Hit evalue** | **Best Hit ARO** | **Best Hit ARO category** |
|---|---|---|---|---|---|
| 5011587 | 5014898 | + | 0 | Nocardia rifampin resistant beta-subunit of RNA polymerase (rpoB2) | determinant of rifamycin resistance - antibiotic resistant gene variant or mutant - antibiotic target replacement protein |
| 3258724 | 3262026 | + | 9.10E-209 | MexF | efflux pump complex or subunit conferring antibiotic resistance |
| 1854615 | 1858097 | + | 3.90E-186 | mfd | antibiotic target protection protein - determinant of fluoroquinolone resistance |
| 306447 | 308336 | + | 2.10E-165 | TaeA | efflux pump complex or subunit conferring antibiotic resistance |
| 3946628 | 3947857 | - | 1.70E-155 | Streptomyces cinnamoneus EF-Tu mutants conferring resistance to elfamycin | gene involved in self-resistance to antibiotic - antibiotic resistant gene variant or mutant - determinant of elfamycin resistance |
| 4428676 | 4430475 | + | 3.00E-137 | Streptomyces rishiriensis parY mutant conferring resistance to aminocoumarin | gene involved in self-resistance to antibiotic - determinant of aminocoumarin resistance - antibiotic resistant gene variant or mutant |
| 3374014 | 3377079 | + | 1.10E-120 | MuxC | efflux pump complex or subunit conferring antibiotic resistance |
| 5764662 | 5765987 | - | 6.10E-95 | Mycobacterium tuberculosis intrinsic murA conferring resistance to fosfomycin | antibiotic resistant gene variant or mutant - determinant of fosfomycin resistance |
| 2848512 | 2850239 | - | 7.20E-88 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 2102806 | 2105910 | + | 4.10E-86 | MuxB | efflux pump complex or subunit conferring antibiotic resistance |
| 5022007 | 5024976 | + | 4.80E-84 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 3948010 | 3950085 | - | 4.90E-83 | tetT | antibiotic target protection protein - determinant of tetracycline resistance |
| 2048562 | 2049896 | + | 4.50E-82 | DnaA | antibiotic target protection protein - determinant of rifamycin resistance |
| 966466 | 969324 | - | 6.30E-81 | Bifidobacteria intrinsic ileS conferring resistance to mupirocin | determinant of mupirocin resistance - antibiotic resistant gene variant or mutant |
| 5445510 | 5447525 | - | 6.00E-78 | msbA | efflux pump complex or subunit conferring antibiotic resistance |
| 1204425 | 1206167 | + | 2.00E-74 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 2899058 | 2900446 | - | 9.20E-70 | patA | efflux pump complex or subunit conferring antibiotic resistance |
| 3653458 | 3655272 | - | 6.60E-68 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 5333013 | 5335559 | + | 6.00E-67 | mfd | antibiotic target protection protein - determinant of fluoroquinolone resistance |
| 151546 | 152436 | - | 1.30E-61 | antibiotic resistant fabI | antibiotic resistant gene variant or mutant - determinant of isoniazid resistance - determinant of triclosan resistance |
| 2832133 | 2833350 | - | 4.00E-61 | macB | efflux pump complex or subunit conferring antibiotic resistance |
| 1412067 | 1413350 | - | 7.20E-61 | patA | efflux pump complex or subunit conferring antibiotic resistance |
| 2215320 | 2216378 | - | 1.10E-59 | vanG | determinant of resistance to glycopeptide antibiotics - protein(s) conferring antibiotic resistance via molecular bypass - antibiotic resistance gene cluster |

| | | | | | |
|---|---|---|---|---|---|
| 783219 | 784238 | + | 1.80E-59 | PmrF | determinant of polymyxin resistance - gene altering cell wall charge |
| 1486617 | 1487831 | + | 2.10E-57 | macB | efflux pump complex or subunit conferring antibiotic resistance |

Table S3: LMECYA167 combined data alignment scaffold RGI tool hits.

| Start position (bp) | Stop position (bp) | Orientation | Best Hit evalue | Best Hit ARO | Best Hit ARO category |
|---|---|---|---|---|---|
| 5011587 | 5014898 | + | 0 | Nocardia rifampin resistant beta-subunit of RNA polymerase (rpoB2) | determinant of rifamycin resistance - antibiotic resistant gene variant or mutant - antibiotic target replacement protein |
| 3258724 | 3262026 | + | 4.50E-208 | MexF | efflux pump complex or subunit conferring antibiotic resistance |
| 1854615 | 1858097 | + | 2.30E-186 | mfd | antibiotic target protection protein - determinant of fluoroquinolone resistance |
| 306447 | 308336 | + | 1.90E-166 | TaeA | efflux pump complex or subunit conferring antibiotic resistance |
| 4428553 | 4430475 | + | 1.60E-152 | Streptomyces rishiriensis parY mutant conferring resistance to aminocoumarin | gene involved in self-resistance to antibiotic - determinant of aminocoumarin resistance - antibiotic resistant gene variant or mutant |
| 3374014 | 3377079 | + | 1.20E-117 | MuxC | efflux pump complex or subunit conferring antibiotic resistance |
| 5764668 | 5765987 | - | 3.00E-94 | Mycobacterium tuberculosis intrinsic murA conferring resistance to fosfomycin | antibiotic resistant gene variant or mutant - determinant of fosfomycin resistance |
| 2848512 | 2850239 | - | 8.40E-89 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 2102806 | 2105910 | + | 4.10E-86 | MuxB | efflux pump complex or subunit conferring antibiotic resistance |
| 5022007 | 5024976 | + | 1.80E-83 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 2048562 | 2049896 | + | 1.00E-81 | DnaA | antibiotic target protection protein - determinant of rifamycin resistance |

| | | | | | |
|---|---|---|---|---|---|
| 966466 | 969324 | - | 2.80E-81 | Bifidobacteria intrinsic ileS conferring resistance to mupirocin | determinant of mupirocin resistance - antibiotic resistant gene variant or mutant |
| 5445510 | 5447525 | - | 3.20E-79 | msbA | efflux pump complex or subunit conferring antibiotic resistance |
| 1288412 | 1290226 | + | 6.40E-79 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 1204425 | 1206167 | + | 1.30E-76 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 3946628 | 3947191 | - | 9.20E-69 | Streptomyces cinnamoneus EF-Tu mutants conferring resistance to elfamycin | gene involved in self-resistance to antibiotic - antibiotic resistant gene variant or mutant - determinant of elfamycin resistance |
| 2899058 | 2900446 | - | 1.00E-68 | patA | efflux pump complex or subunit conferring antibiotic resistance |
| 5333955 | 5335481 | + | 9.50E-68 | mfd | antibiotic target protection protein - determinant of fluoroquinolone resistance |
| 3653458 | 3655272 | - | 1.90E-67 | sav1866 | efflux pump complex or subunit conferring antibiotic resistance |
| 151546 | 152436 | - | 1.30E-61 | antibiotic resistant fabI | antibiotic resistant gene variant or mutant - determinant of isoniazid resistance - determinant of triclosan resistance |
| 1412067 | 1413350 | - | 7.20E-61 | patA | efflux pump complex or subunit conferring antibiotic resistance |
| 783219 | 784238 | + | 2.90E-60 | PmrF | determinant of polymyxin resistance - gene altering cell wall charge |
| 2832133 | 2833350 | - | 5.80E-60 | macB | efflux pump complex or subunit conferring antibiotic resistance |
| 1486617 | 1487831 | + | 5.40E-58 | macB | efflux pump complex or subunit conferring antibiotic resistance |

| | | | | | |
|---|---|---|---|---|---|
| 222381 | 224822 | + | 3.90E-55 | rphB | determinant of rifamycin resistance - antibiotic inactivation enzyme |