Novel scalable approaches for the computational analysis of bacterial genomes

Cumulative inaugural dissertation
for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

by

Oliver Schwengers

submitted to the
Faculty of Biology and Chemistry (FB08)
prepared at the
Department for Bioinformatics & Systems Biology

Justus Liebig University Giessen
Giessen 2021

First reviewer: Prof. Dr. Alexander Goesmann, Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany

Second reviewer: Prof. Dr. Trinad Chakraborty, Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, Germany

Contents

1 Abstract

Over the last decades, the giant progress of DNA sequencing led to increased throughput and tremendously reduced costs resulting in a broad accessibility and applicability of these technologies and thus revolutionized the entire field of microbial genomics. Today, these developments allow the sequencing of large groups and entire cohorts of bacterial genomes in a timely manner, whereas a mere decade ago, this was only feasible for a few single genomes. Now, hundreds of thousands of sequenced bacterial genomes are available in public databases and vast numbers of genomes are sequenced worldwide on a daily basis without any foreseeable climax. Many fields of research benefit from these developments, in particular medical microbiology and epidemiology. Hence, genome-based analyses have nowadays become essential tools for the detection, classification, typing and comparison of special-interest genes and pathogenic genomes at various levels. At the same time, IT is revolutionized alike by new developments like cloud computing and software containerization techniques. Modern software engineering paradigms and frameworks have recently emerged and provide new opportunities for scalable computations on distributed and heterogeneous infrastructures that in turn imply new technical premises. Albeit the mere sequencing of bacterial genomes as well as computing capacity in general are not the major limiting factors anymore, the comprehensive, timely and standardized analysis of large bacterial whole-genome sequencing data however remains an issue of rising importance.

Therefore, it was the aim of this thesis to address these challenges and provide novel bioinformatic approaches and software tools for the scalable high-throughput analysis of whole-genome sequencing data of large bacterial cohorts. An automated and comprehensive workflow was designed and implemented in a portable, scalable and user-friendly software tool **ASA³P**. It supports data from all contemporary DNA sequencing platforms conducting the streamlined processing and analysis from raw reads to assembled, annotated and comprehensively characterized genomes including comparative analyses. The software provides both vertical and horizontal scalability allowing researchers to take advantage of distributed and versatile computing infrastructures. Results are presented as integrated, human-readable and interactive reports. Two further contributions address issues that have arisen from the design of this workflow. For the integrated analysis of plasmids, a novel methodology has been developed for the automated and taxonomy-independent detection and characterization of plasmid-borne contigs from fragmented bacterial draft assemblies. As a new approach to this problem, the natural distribution bias of protein-coding gene families among chromosomes and plasmids is utilized, which achieves a robust and competitive classification performance. This new methodology was implemented in the software tool **Platon**, which also provides additional plasmid characterizations. A third contribution addresses the robust and accurate but rapid computation of mutual genome distances that is required for the automated selection of high-quality reference genomes and whole-genome-based taxonomic classifications. As the large amount of available genome sequences poses increasing hurdles to these steps in terms of data accessibility, performance and runtimes, a new software tool called **ReferenceSeeker** combining existing methodologies was developed and complemented by the provisioning of integrated and customizable databases. Noteworthy, its application is not limited to microbial genomes alone, but DNA sequences in general, including plasmids.

These three bioinformatics solutions have been used in various published and unpublished studies and proven as useful software tools for researchers in the field of medical microbiology. In particular, ASA³P enables researchers to take advantage of modern and scalable IT resources and provides access to a diverse set of proven bioinformatics software tools. Hence, even more bacterial genomes and larger cohorts thereof can be processed, characterized and compared among each other, allowing researchers to keep pace with DNA sequencing technologies and future demands. Due to its extensible framework, the application of ASA³P is however not restricted to medical microbiology applications, but can be expanded and adapted to applications within the much larger field of microbial genomics. Furthermore, several ideas for further improvements and potential new software solutions emerged from this work that opened new research questions and established interesting subjects for future investigations.

Zusammenfassung

Immense Fortschritte auf dem Gebiet der DNA-Sequenzierung führten in den letzten Jahrzehnten zu einer enormen Kostenreduzierung und Erhöhung des weltweiten Sequenzieraufkommens. Die damit einhergehende weite Verbreitung und einfache Anwendbarkeit dieser Technologien revolutionierte in Folge umfassend das gesamte Gebiet der mikrobiellen Genomik. Noch vor einem Jahrzehnt undenkbar, ist es heute möglich, zeitnah große Kohorten ganzer Bakteriengenome zu sequenzieren. Öffentliche Datenbanken bieten heutzutage Zugang zu hunderttausenden Bakteriengenomen und täglich kommen ohne erkennbare Verlangsamung unzählige hinzu. Von diesen Entwicklungen profitieren viele Forschungsgebiete, insbesondere die medizinische Mikrobiologie und Epidemiologie. Computergestützte genetische Analysen sind zu unverzichtbaren Werkzeugen für den Nachweis, die Klassifizierung, Typisierung und den Vergleich pathogener Genome auf unterschiedlichsten Ebenen geworden. Gleichzeitig erhielten neue Entwicklungen wie Cloud Computing und Softwarecontainerisierung Einzug in die Informationstechnologie und revolutionieren diese gleichermaßen. Moderne Frameworks und Software-Engineering-Paradigmen bieten neue Möglichkeiten für skalierbare Berechnungen auf verteilten und heterogenen Infrastrukturen, welche jedoch neue technische Ansätze und softwareseitige Anforderungen voraussetzen. Auch wenn die Sequenzierung bakterieller Genome sowie notwendige Rechenkapazitäten zur Analyse keine wesentlichen limitierenden Faktoren mehr darstellen, ist die zeitnahe, eingehende und standardisierte Analyse großer Kohorten bakterieller Genomsequenzierungsdaten gleichwohl Gegenstand aktueller bioinformatischer Forschung.

Ziel dieser Arbeit war es daher, diese Herausforderungen zu adressieren und neue bioinformatische Ansätze und Softwaretools für die skalierbare Hochdurchsatzdatenanalyse von Gesamtgenomsequenzierungen großer bakterieller Kohorten zu entwickeln. Hierzu wurde ein automatisierter und umfassender Workflow entworfen und in dem portablen, skalierbaren sowie benutzerfreundlichen Softwaretool **ASA^{3}P** implementiert. Dies unterstützt alle verbreiteten DNA-Sequenzierungsplattformen sowie die automatische Prozessierung und Analyse der Daten hin zu assemblierten und annotierten Genomen mit anschließender umfangreicher Genomcharakterisierung und komparativen Analysen aller Genome einer Kohorte. Die portable Software bietet eine sowohl vertikale als auch horizontale Skalierbarkeit, welche es Forschenden ermöglicht, verteilte und vielseitige

Computerinfrastrukturen zu nutzen. Alle Ergebnisse werden in standardisierten bioinformatischen Dateiformaten ausgegeben sowie als integrierte, für Menschen lesbare, interaktive Berichte präsentiert. Aus der Gestaltung dieses Workflows ergaben sich neue Fragestellungen, welche in zwei weiteren Beiträgen dieser Arbeit behandelt wurden. Für die integrierte Analyse von Plasmiden wurde eine neuartige Methodik für den automatisierten und taxonomieunabhängigen Nachweis plasmidärer Contigs aus bakteriellen Draftassemblierungen mit anschließender Charakterisierung entwickelt. Als neuer Ansatz zu diesem Problem werden hierzu Unterschiede in der natürlichen Verteilung proteinkodierender Genfamilien zwischen Chromosomen und Plasmiden genutzt, wodurch eine robuste und kompetitive Klassifizierung erreicht wird. Diese neue Methodik wurde mitsamt umfangreicher Plasmidcharakterisierungen in dem Softwaretool **Platon** implementiert. Ein dritter Beitrag adressiert die schnelle und genaue Berechnung bidirektionaler Genomdistanzen, welche für die automatisierte Auswahl hochqualitativer Referenzgenome und gesamtgenombasierter taxonomischer Klassifikationen erforderlich ist. Die schiere Menge verfügbarer Genomsequenzen stellt jedoch ein immer größer werdendes Hemmnis für diesen Auswahlprozess bezüglich Datenverfügbarkeit, Qualität und Laufzeit dar. Dazu wurde ein neues Softwaretool namens **ReferenceSeeker** entwickelt, welches bestehende Methoden kombiniert und durch die Bereitstellung integrierter und erweiterbarer Datenbanken ergänzt wurde. Ein wichtiger Vorteil der Software ist dessen breite mikrobielle Anwendbarkeit, welche nicht auf Bakteriengenome beschränkt ist, und darüber hinaus auch allgemeine DNA-Sequenzen, insbesondere Plasmide umfasst.

Diese neuen bioinformatischen Softwarelösungen wurden in verschiedenen publizierten Studien verwendet und haben sich als nützliche Werkzeuge für Forschende auf dem Gebiet der medizinischen Mikrobiologie bewährt. Insbesondere ermöglicht ASA³P die Vorteile moderner und skalierbarer IT-Ressourcen zu nutzen, und bietet Zugang zu einer Vielzahl bewährter bioinformatischer Softwaretools und Datenbanken. So können immer mehr Bakteriengenome und größere Kohorten verarbeitet, charakterisiert und miteinander verglichen werden, und Forschende mit zukünftigen Anforderungen Schritt halten. Aufgrund seines modularen Designs ist die Anwendung von ASA³P jedoch nicht auf Anwendungen in der medizinischen Mikrobiologie beschränkt, sondern kann erweitert und an vielfältige Anwendungen innerhalb der mikrobiellen Genomik angepasst werden. Des Weiteren gingen aus dieser Arbeit zahlreiche Ideen für weitere Verbesserungen und potenzielle neue Softwarelösungen hervor, welche neue Forschungsfragen aufwerfen und interessante Themen für zukünftige Untersuchungen bieten.

2 Introduction

"… there are 100 million times as many
bacteria in the oceans (13 × 10^{28})
as there are stars in the known universe."

Microbiology by numbers
Nature Reviews Microbiology, 2011

2.1 Rationale and outline

Microbes are the oldest organisms on earth and preceded all other living beings, especially multicellular eukaryotes like animals and plants, by nearly three billion years [1]. They have been the pioneers of this planet and the foundation of the biosphere, from both an evolutionary as well as an environmental perspective [2]. At all times throughout history, humans have lived in complex ecosystems and ambivalent relationships with these microorganisms. On the one hand, they populate complex ecological niches on the surface of and within multicellular eukaryotes like plants and humans [3]. For a long time, the endogenous human flora was poorly understood [4,5], but step by step more and more white spots on this map have been erased. Today, we know that the number of commensal bacteria colonizing the human body approximately equals or even exceeds the number of human cells [6,7]. Only recently, we started to grasp that these diverse communities pose an essential natural line of defense against pathogenic microorganisms and therefore play an important role for human health [8–11]. On the other hand, there is a large number of well-known pathogenic bacteria causing severe infectious diseases. For millennia, mankind has been severely threatened by many of these, which had a tremendous impact on the human population on a global scale. For example, the plague caused by *Yersinia pestis* [12] has been accountable for many large historical outbreaks [13,14]. For instance, the medieval pandemic, which had tremendous socio-economic effects, has been described as "one of the most dramatic examples ever" [15]. Although the plague is deemed vanquished, still today, there are occasional local outbreaks in different parts of the world [16,17]. However, besides these small and large-scale epidemics, many infectious diseases are of a rather permanent nature posing a lasting burden for humankind. For example, the typhoid fever caused by *Salmonella enterica* serovar Typhi [18] led to estimated 21.7 million infections resulting in 217,000 deaths in the year 2000 alone [19,20]; for tuberculosis caused by *Mycobacterium tuberculosis* [21], 10 million infections and nearly 1.2 million attributable deaths were reported for 2019 [22]; and for cholera caused by *Vibrio cholerae,* 1.2 million cases were reported in 2017 [23]. To this brief exemplary list, a large set of pathogens must be added that caused 600 million food-born illnesses in the year 2010 – amongst these the most severe agents, *e.g. Escherichia coli*, *Campylobacter*, *Listeria monocytogenes* and *Salmonella* as well as pneumonia-causing bacteria, *e.g. Streptococcus pneumoniae* and *Haemophilus influenzae* type b [24].

In the mid-20th century, the human quest for biological survival of infectious diseases was deemed as succeeded due to the discovery, and shortly afterwards, broad availability of antibiotics. However, over the course of the last decades, this evolutionary competitive edge constantly lost its effectiveness as more and more bacteria have become resistant to many commonly used antibiotic drugs. Sadly, humankind has induced and is rapidly approaching a situation in which bacterial infections could become untreatable again. Unfortunately, even resistances against so-called last-resort antibiotics are detected more frequently, as for instance, the *mcr-1* gene conferring resistance against colistin [25–28]. These developments have evolved to a severe public health issue and a threat for people worldwide. Already in 1990, the Nobel Prize winner Joshua Lederberg stated: “We live in evolutionary competition with microbes – bacteria and viruses. There is no guarantee that we will be the survivors” [29]. The magnitude of this global medical crisis has become distressingly clear in a recent study estimating that, without effective countermeasures, about 10 million people could die annually in 2050 due to antibiotic-resistant bacteria [30]. Likewise, it has become clear that the implied economic burden of antimicrobial resistance (AMR) treatments is massive [31,32].

Fortunately, simultaneously to the emergence of these threats, constant and steep scientific progress in the fields of molecular biology, microbiology, medicine and bioinformatics facilitated new methodologies for deeper investigations of the microbial universe as well as for the surveillance and outbreak detection of human pathogens [33,34]. The advent of next-generation sequencing (NGS) technologies commenced a new era of high-throughput DNA sequencing in which bacterial genomes are investigated in hitherto unknown resolution and unprecedented numbers. New genome-based approaches have expeditiously become routine for the effective surveillance and precise tracing of infection chains within pathogen outbreaks [35]. Furthermore, whole-genome sequencing (WGS) based antibiotic susceptibility testing (AST) matures and might replace molecular-based phenotypic AST in the medium term [36]. Potential future applications might comprise real-time on-site or even point-of-care sequencing of clinical samples providing instant and actionable results. According to these global developments, the described threat of antibiotic-resistant bacteria is on the agenda of policy makers and health professionals worldwide. For instance, the European Union installed a union-wide surveillance system led by the European Centre for Disease Prevention and Control (ECDC) to collect, analyze and report data on antibiotic-resistant bacteria through a network of national surveillance systems in which all member states participated [37].

This strong rise of antibiotic resistance-related DNA sequencing projects in both academia and public health authorities has contributed to the genesis of a large number of bioinformatics software tools and databases. Although there is an obvious trend in implementing open data standards, as for instance the well-known FAIR (findability, accessibility, interoperability and reusability) principles [38], only few if any open standards in clinical WGS are in place. Even worse, this large number of available bioinformatics solutions fostered an obvious lack of consensus regarding the choice of optimal methodologies, algorithms, software implementations and databases, which often need to be combined in complex workflows exacerbating these issues [39]. However, to fully exploit the vast potential of these promising scientific opportunities, the demand for standard operating procedures, common application interfaces and analysis workflows needs to be addressed in order to handle the implied overwhelming complexity.

In addition to the expanding complexity of data analysis workflows, the sheer amount of existing data is increasing by magnitudes and thus challenges established information technology (IT) infrastructures. In 2015, it has been estimated that the yearly acquisition of DNA sequencing raw data could rise to a worldwide level of one zettabyte – a trillion gigabytes (GB) – in 2025 [40]. This mind-boggling amount of raw data will push IT requirements to unknown levels. Data analysis solutions provided by dedicated and specialized but centralized online services might soon be outpaced by rising sequencing outputs and their usage will be limited by the heavily used networks connecting local DNA sequencing facilities with centralized online services. These bottlenecks will put a strong pressure on the way this raw data will be processed and finally foster new approaches: away from centralized online services and towards local computers on the one hand and nearby scalable cloud computing infrastructures (CCIs) on the other.

Due to the increasing power of standard consumer hardware, it is nowadays still feasible to analyze small data sets on local computers in a timely manner avoiding the limiting public transportation of raw data and limited computing capacities of centralized software solutions. However, increasing amounts of data like, for example, combinations of novel DNA sequencing technologies and growing cohorts of large numbers of samples, will inevitably require computing resources beyond the capabilities of local computers and centralized online services. In this context, distributed CCIs provide several advantages. First and most importantly, flexibility: CCIs are able to dynamically and rapidly provide vast computing resources on demand and thus build the technical foundation for the timely analysis of medium to even very large data sets. For example, suitable amounts of computing resources can be requested and instantly provided according to data sizes at hand and requested analysis workflows. Furthermore, CCIs provide the opportunity and technical solutions to dynamically adapt provisioned computing resources to changing requirements over the course of multistep data analysis workflows. Hence and second costs: costly upfront expenditures for local computing infrastructures can be spared. Furthermore, CCIs are able to take advantage of economies of scale by using shared pools of computing resources thus achieving much higher overall usage statistics compared to potentially underused local computing infrastructures. These advantages are complemented by a considerable potential to improve on operating expenses in terms of power efficiency. Hence, CCIs provide huge potentials for economic but also ecological cost reductions. Third data throughput: compared to centralized online services, distributed CCIs are able to provide on average closer hosting sites. Thus, the physical distance and the number of network endpoints that the data must pass on its way from DNA sequencing to data processing facilities is potentially reduced. This might result in reduced overall network usage and shorter data transportation periods.

Finally, information gained from raw data processing and analysis, either locally or CCI-based, could then be submitted to dedicated centralized online services and information repositories running sophisticated information aggregation and big data algorithms. This higher level information could then be analyzed by specialized software solutions potentially exploiting artificial intelligence to create new knowledge from this plethora of information. However, in order to keep pace with these developments, bioinformatics software tools will need to fulfill an increasing number of technical requirements to play their role in a growing stack of decentralized analysis workflows: they (i) need to be both easily portable and installable for non-IT experts via common software environment management systems, *e.g.* Docker [41], Podman [42] or BioConda [43], (ii) require implementations that allow vertical and preferably horizontal scalability to different work loads, (iii) need to follow community best practices in terms of standardized interfaces and file formats allowing the seamless interconnection with other tools and the integration into larger workflows.

To address these issues, novel approaches and bioinformatics software solutions have been developed and are described in this thesis for the automated high-throughput analysis of bacterial WGS data from single genomes and larger cohorts thereof. Three new bioinformatics software tools and a new methodology are described as scientific contributions to the field of microbial genomics. The main contribution comprises the design of an automated analysis workflow for the processing and comprehensive characterization of cohorts of sequenced bacterial genomes as well as its implementation fulfilling the described contemporary requirements. Two further contributions resulted from the design of the aforementioned workflow: the development of a taxonomy-independent new methodology for the fully automated and robust detection and characterization of plasmid-borne contigs resulting from bacterial draft assemblies, and the rapid but accurate determination of suitable reference genomes for the automated selection and usage in reference-based analysis workflows.

The following section 2.2 provides a brief historical recapitulation and introduction into the field of medical microbiology from its beginning to the most recent developments regarding outbreak detection, surveillance, investigation and characterization of bacterial organisms as a major contemporary driving force of the global demand for sequenced sets of bacterial genomes. This is followed by an introduction to the field of DNA sequencing and related developments in downstream bioinformatics analysis in section 2.3. To describe these immense breakthroughs and to highlight the increasingly steep acceleration of scientific progress in these fields, a rather large time frame was chosen. As one key driver for the large-scale sequencing of bacterial isolates, the global dissemination of antibiotic-resistant bacteria as well as the implied emerging threats are described in section 2.4. This is followed by section 2.5 introducing the *in silico* analysis of bacterial plasmids as a key vehicle for the global dissemination of antibiotic resistance genes. Section 2.6 describes applications for reference genomes and challenges regarding the optimal selection thereof. Section 2.7 provides a short layout of contemporary challenges in microbial bioinformatics for the medical microbiology of the coming 2020s. The introduction is concluded with a description of the scientific challenges and gaps in the field of microbial bioinformatics regarding the automated high-throughput analysis of bacterial WGS data in section 2.8. After short summaries of the peer-reviewed publications being part of this thesis, these contributions are described and discussed in the light of the explicated scientific background.

2.2 Medical microbiology and epidemiology in the course of time

The second half of the 19th century was the natal hour of medical microbiology as we know it today. Robert Koch, Louis Pasteur and Paul Ehrlich, just to name a few besides many prominent medical and microbiological scientists, made groundbreaking discoveries that revolutionized our fundamental understanding of infectious diseases and the way we treat them today. Until that time, it had been believed that infectious diseases were caused by polluted air arising from decayed organic matter, condensed in the theory of miasma [44]. Originating from religious thoughts, Hippokrates formed a naturalistic view of the origin of infectious diseases, which inhabited European cultures for more than 2,000 years until the mid-19th century [45–47]. In 1854, a severe cholera outbreak in London’s Soho district caused about 644 deaths [48]. In the middle of this outbreak, the young British physician John Snow doubted the miasma theory in favor of a waterborne transmission [49]. Although Snow achieved to find the spatial origin of the outbreak by combining death cases with geographical data, he was not able to prove his ideas. Posthum, his conception was endorsed by the medical statistician William Farr, a former vital supporter of the miasma theory, by the statistical analysis of death rates [50]. In the course of the following decades, the miasma theory more and more struggled and was finally taken over by the modern *germ theory*. This opposing theory, conveying the conception that microbes are the cause of infections, recurrently came up and was described in basic versions throughout the ages, from Ibn Sina in 1025 [51] to Girolamo Fracastoro in the mid-16th century [52]. But in 1876, this theory was finally proven by the description of the full lifecycle of *Bacillus anthracis* – maybe the most significant breakthrough in the field of medical microbiology. For the first time in history, the full progression of an infectious disease could be described in combination with and in the light of the causing agent that has bedeviled mankind ever since [53]. Only six years later, in 1882, Robert Koch published the discovery of *Mycobacterium tuberculosis*, the agent and cause of one of the most deadly and feared infectious diseases causing one out of seven deaths in Europe in these times [21]. A mere year after, Koch discovered the cholera agent *Vibrio cholerae* and thus finally proved John Snow right [54]. Several of these findings have been condensed and summed up in his famous postulates, which are still in use today and led to the discovery and description of many bacterial pathogens in the following years [55]. These major discoveries, among many others in the “golden age” of bacteriology, laid the cornerstone for numerous following scientific discoveries in molecular biology and genetics and thus built the fundamental scientific basis for our today's understanding of medical microbiology and epidemiology.

Over the course of the following century, bacterial species were described in profusion, hugely expanding the known microbial world. In 1980, a large reorganization of the bacterial nomenclature reduced redundancy resulting in about 2,300 unique taxonomic names [56]. In order to distinguish and finally identify known and new bacterial species, an elaborate and time-consuming phenotypic description had to be conducted, comprising the morphological description and stainings by visual inspections as well as metabolic classifications based on chemical assays. Not until the discovery and understanding of the DNA as the encoding molecule of the genetic information in the middle of the 19th century, it was possible to take advantage of the inherent ultimate blueprint of all living beings as a new approach for the description of and discrimination between bacterial species. This discovery along with concordant progresses in molecular biology started a shift from mere morphological and macromolecular, *i.e.* phenotypic, descriptions to modern DNA-based, *i.e.* genotypic, descriptions and characterizations of bacterial organisms. Based on these genotypic methods, the definition of bacterial species, the typing of closely related groups and even the identification of single strains became conductible leading to revolutionary developments in medical microbiology as well as new tools for modern epidemiology.

These advances in the use of molecular markers and genetics posed an epidemiological game changer. For instance, in the 1970s DNA-DNA hybridization was introduced as a bilateral method to differentiate between bacterial species. Genomes that showed more than 70% DNA-DNA homology under given conditions were considered to belong to the same species [57]. Improving from mere species delineation, in 1984 Schwartz and Cantor developed the pulsed-field gel electrophoresis (PFGE) as an improvement to normal gel electrophoresis [58]. This new technique enabled the DNA-based fingerprinting of bacterial organisms with previously unreached resolution. Now, this fine-grained discrimination of bacterial strains allowed for the tracing of individual strains down an epidemiological chain. In 1996, the US Centers for Disease Control and Prevention (CDC) started a programme named PulseNet aiming to create a large compilation of PFGE-based DNA-fingerprints of foodborne bacterial human pathogens [59]. As soon as DNA-fingerprints of new clinical isolates from hospitalized patients were available, they could be compared to those stored in the database. Matches helped to subsequently build a reliable link between a patient and the food or its production environment helping to frame foodborne pathogen outbreaks. Convinced by the results of this project, until 2001, all US public health state laboratories engaged into PulseNet. Until 2015, this database had grown to one million records covering more than 500,000 DNA fingerprints of *Salmonella* genomes alone and nearly 90,000 patterns were queried and compared from participating public health authorities [60]. By exploiting the pathogens' genomes in order to identify and type them and finally trace outbreaks, PulseNet became an epidemiological story of success.

Meanwhile, a new approach called multilocus sequence typing (MLST) was proposed in 1998. Instead of comparing physical fingerprints of DNA fragments, this new methodology was based on recently emerging DNA sequencing technologies. Exploiting species specific expert knowledge, DNA sequences of a tiny fraction of housekeeping genes are collected and assigned arbitrary numbers. Combinations of these allele numbers are assigned numbers again resulting in a digital fingerprint that is simple to use, share and communicate [61]. Advantages in terms of electronic portability and much higher strain resolution instantly led to the creation of MLST schemes for many pathogenic species [62–68] as well as bioinformatics software tools and data sharing platforms for the analysis and sharing of pathogenic bacterial genome sequence types [69,70].

However, in 2005, when the first commercial NGS platforms entered the market, it became clear that these new techniques had the potential to revolutionize microbiology and epidemiology, again [71]. Only five years later, the PulseNet project used a WGS approach to investigate an ongoing outbreak of *Vibrio cholerae* in Haiti with 93,000 cases and 2,100 attributable deaths [59,72]. Shortly after, in 2013, the CDC started to use WGS techniques for the routine surveillance of *Listeria*. By doing so, more pathogen genome clusters were detected, and more outbreaks could be solved than ever before. Phylogenetic investigations exploiting the high genetic resolution of WGS approaches were shown to be in line with epidemiological data, thus helping to reconstruct outbreaks. Hence, WGS has transformed the surveillance of *Listeria* related outbreaks [73].

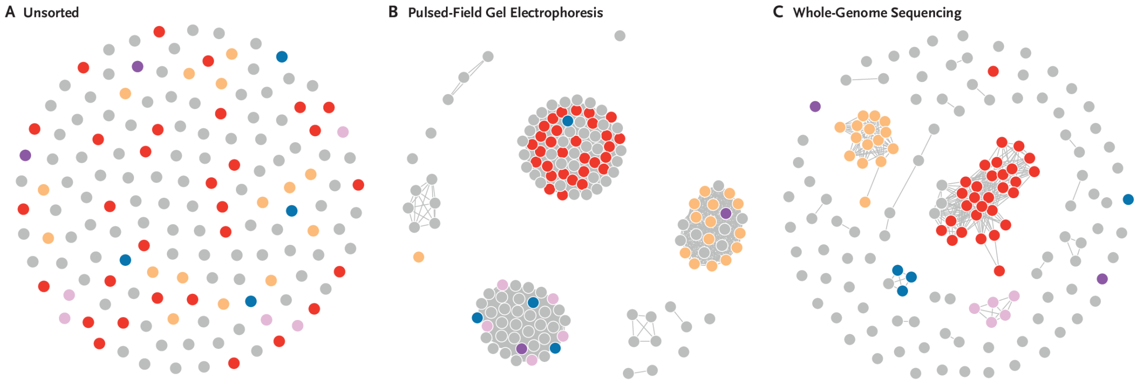


Figure 1: Exemplary depiction of cluster resolution levels of different molecular methods for outbreak detections and investigations.

Isolates of *Salmonella enterica* sampled during an outbreak in the USA in 2018 have been investigated and clustered using different genomic methods. Individual dots represent cases of gastroenteritis with sampled and investigated isolates. Gray dots represent cases that were determined not to be linked to an outbreak; colored dots represent cases linked to confirmed outbreaks. A) For demonstration purposes, cases are randomly placed representing unclustered samples if no data is available. B) Clustering of isolates according to PFGE. C) Clustering of isolates according to WGS methods. Reprinted with permission from The NEW ENGLAND JOURNAL of MEDICINE [77], Copyright © 2019, Massachusetts Medical Society.

This unprecedented resolution of WGS methods, combined with publicly available genome sequences allowed the exact and rapid typing and identification of bacterial pathogens on a global scale and hence, quickly transformed the way how outbreaks were investigated in general [74,75]. It could be shown that the fine-grained resolution of WGS approaches clearly outperform former methodologies like serotyping, PFGE and MLST [76] and that this increased genomic resolution is highly advantageous for epidemiological investigation of outbreak clusters and potential transmission routes (Figure 1) [77]. Soon, it became a routine standard for the surveillance of many foodborne bacterial pathogens in 2016 [59]. Furthermore, a contemporary study came to the conservative estimation that PulseNet helped to avoid about 266,000 illnesses from *Salmonella* and nearly 10,000 illnesses from *Escherichia coli* annually, thus reducing medical and productivity costs by about US$500 million [78]. Spurred by these large medical successes, modern WGS techniques became the new PulseNet gold standard for foodborne pathogen surveillance [79]. Furthermore, roughly 20 years after its inauguration, the PulseNet project was expanded globally fostering new international collaboration and standardization [80]. Against the background of its advantages over preceding technologies, it is not surprising that WGS was rapidly picked up and has successfully been used in many other scientific and public health projects [81,82]. For instance, the GenomeTrakr project was the first distributed network of US state, local, federal as well as international laboratories solely applying WGS based approaches for the surveillance of foodborne pathogens [83,84]. By the end of 2019, nearly 500,000 bacterial genomes were sequenced and stored publicly along with equally important metadata. It goes without saying that PulseNet and GenomeTrakr are accompanied with similar sequencing projects and databases all over the world. Further examples of contemporary pathogen genome platforms for epidemiology are EnteroBase [85] and Pathogenwatch [86]. The latter uses pre-assembled bacterial genomes to focus on the subtyping of isolates, prediction of antibiotic resistances and the subsequent comprehensive interactive visualization thereof combined with related metadata.

Of course, the nowadays broad application of WGS is neither limited to the analysis of bacterial pathogens nor to the detection and surveillance of pathogenic outbreaks, alone. DNA sequencing of large numbers of bacterial isolates and cohorts of closely related genomes from various sources, *e.g.* different hosts and environments, has tremendously contributed to our current understanding of bacterial life. Besides the general organization of the genome itself, this also comprises fundamental underlying genetic mechanisms, different sizes of pan-genomes *i.e.* the entirety of genes within a given population, taxonomic diversity and complexity, genomic population structures and evolutionary dynamics on various scales (Figure 2) [57,87–96]. In regards to medical microbiology, the broader application of WGS undoubtedly propelled a better understanding and deeper knowledge of bacterial pathogenicity [97,98], virulence and host adaptations [99]. Likewise, our understanding of the spread and dissemination of virulence factors and antibiotic resistance genes [100] via horizontal gene transfer highly benefits from comparative studies that take into account more and more genome sequences from either different species or intra-species strains [101].

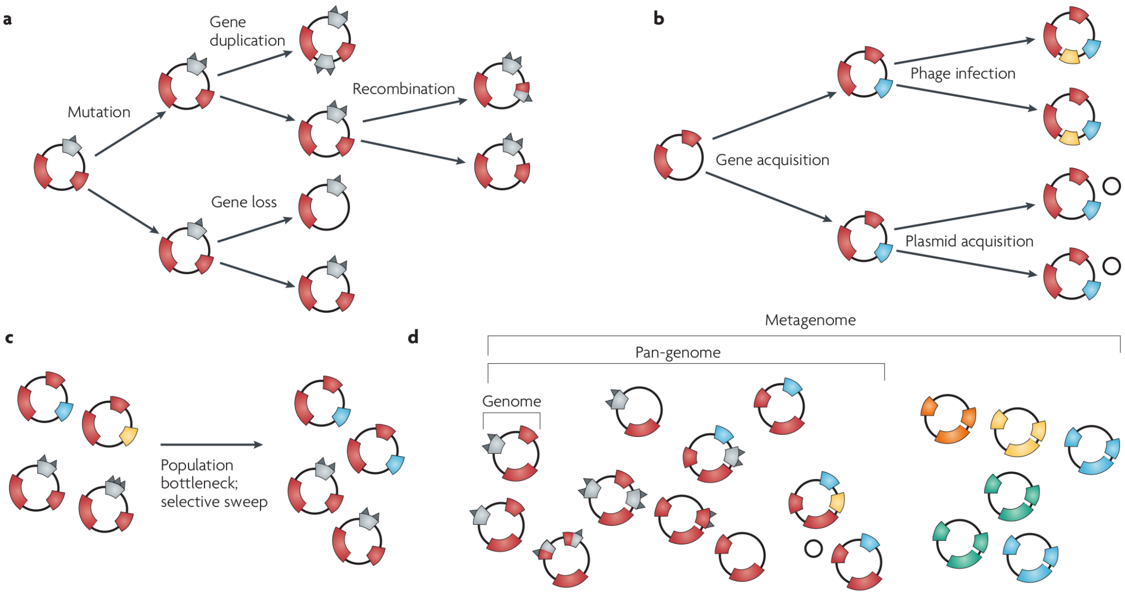


Figure 2: Molecular evolutionary mechanisms propelling genetic diversity and complexity within bacterial populations at different scales.

a) Inherent molecular mechanisms propelling genomic diversity between species and strains. b) Horizontal gene transfer mechanisms driving the exchange of genetic material between species and shaping intra-species population structures. c) Evolutionary selective mechanisms shaping populations. d) Populations at different scales: from a single genome to the pan-genome covering the entirety of all genes of a given species to a metagenome covering all genomes in a given microbial community. Reprinted with permission from Nature Reviews Microbiology [57], Copyright © 2008, Springer Nature.

2.3 The coevolution of DNA sequencing and bioinformatics

"… a knowledge of sequences could contribute much
to our understanding of living matter."

Frederick Sanger
Biographical, 1980

In the first half of the 20th century, several revolutionary experiments made by Griffith, Avery, MacLeod, McCarty, Hershey and Chase [102–104] finally confirmed that DNA was the biological material that stores and transports the genetic information. Due to investigations of the crystallographic structures of DNA created by Franklin and Wilkins, in April 1953 Watson and Crick were able to finally solve its three-dimensional structure [105]. Further discoveries like the *one gene one enzyme* hypothesis [106], the operon concept [107] and finally the deciphering of the genetic code [108] constituted the foundation of modern genetics and all related scientific disciplines, especially present DNA sequencing.

2.3.1 Whole-genome shotgun sequencing

In the mid-1970s two influential protocols for DNA sequencing were published by Sanger and Coulson [109] and Maxam and Gilbert [110]. Only two years later in 1977, Sanger and colleagues achieved to sequence the first entire genome of the bacteriophage ϕX174 [111] and published a new method for DNA sequencing using chain-terminating dideoxy inhibitors [112]. This technique was widely adopted and hence can be considered as the birth of *first generation* DNA sequencing. Three years later, Frederick Sanger was awarded his second Nobel prize for these contributions. At the time of his award ceremony, he claimed his conviction that "… a knowledge of sequences could contribute much to our understanding of living matter." [113] – a well understated claim looking back in retrospect. The first version of a sequenced genome, the bacteriophage ϕX174, had a genome length of 5,375 nucleotides [111]. However, the increasing amount of sequenced DNA fragments and resulting assembly sizes began to challenge their manual editing and paper-based organization, thus raising the demand for computer aided methods. In 1977, Rodger Staden published a first set of computer software tools supporting researchers by the *in silico* storage, editing and analysis of both DNA and amino acid sequences specifically designed “for use by people with little or no computer experience” [114]. Two years later, Staden published a computer-aided DNA sequencing strategy along with an improved version of its software package [115] to deal with the increasing rate of DNA sequencing. In the early 1980s, this rate was further enhanced by new cloning protocols of small and random fragments from DNA restriction enzyme digestion, which led to standardized DNA shotgun protocols [116,117]. Meanwhile, first DNA databases evolved [118] which quickly led to the foundation of two major repositories for DNA submissions shortly after in 1986 whose successors are still actively maintained, today, *i.e.* the GenBank and the European Molecular Biology Laboratory (EMBL) data library [119,120]. In the following years between 1985 and 1990, the groundbreaking chain-terminating technology was further improved [121–125]. Concurrently, the first algorithms for the sensitive and timely sequence similarity searches evolved and first usable implementations were published as for instance FASTA and BLAST (“basic local alignment search tool”) [126–128]. The latter is still in use today using an omnipresent file format standard introduced by FASTA. These developments allowed computer-aided searches for similar sequences in DNA and protein sequence collections of hitherto unprecedented size.

Up to this point, the largest sequenced genome was that of the bacteriophage lambda with a genome size of 48,502 base pairs [129]. The combinatorial complexity of assembling hundreds and thousands of sequenced DNA fragments was a tedious and demanding task that led to the development of many bioinformatics software tools for the automated assembly of such sequenced DNA fragments into larger contigs [130]. However, the progress made in DNA sequencing technologies steeply increased the number of such fragments and likewise the implied computational requirements. This was a severe hurdle that limited WGS projects targeting larger genomes [2]. As a consequence thereof, more efficient assemblers evolved, which allowed the assembly of larger genomes [131,132]. In 1995, the first complete genome of a free living organism, the bacterium *Haemophilus influenzae*, was sequenced [133] followed by the genome of *Mycoplasma genitalium* shortly after [134]. Over the course of the following five years, the genome sequences of about 30 microbes were published [2] comprising many bacterial pathogens, *e.g.* *Mycoplasma pneumoniae* [135], *Escherichia coli* [136], *Bacillus subtilis* [137], *Helicobacter pylori* [138], *Borrelia burgdorferi* [139], *Treponema pallidum* [140] and in 1998 *Mycobacterium tuberculosis* [141] – 116 years after Robert

Koch initially described its pathogenicity. At the end of the 20th century in 1999, for the first time, two genomes of unrelated strains from the same species *Helicobacter pylori* were comprehensively compared [142] by taking advantage of the aforementioned bioinformatics tools FASTA and BLAST [127,128] in order to align and identify orthologous and paralogous genes. In the beginning of the 21st century, in 2001 bacterial genome comparisons were taken to a new level as, for the first time, two entire bacterial genomes were sequenced, annotated and extensively compared against each other in a single scientific publication gaining new insights into the genomic complexity of *Staphylococcus aureus* providing evidences of horizontal gene transfers [143].

In the following years, the large number of sequenced and publicly available microbial genomes posed an enormous fundus for new discoveries leading to new insights into the genetic repertoire and characteristics of microbial genomes. This in turn enabled the development of new algorithms and software tools exploiting this knowledge for the automated prediction of coding and non-coding genes [144–147]. As a consequence, the automated prediction and comparison of microbial, especially bacterial, genes enabled the reconstruction of genetic networks underpinned by the detection of orthologous genes from many genomes. These networks could be further combined with metabolic pathways leading to comprehensive and integrated genome and pathway databases like EcoCyc [148] and KEGG [149]. Furthermore, the growing number of individual smaller sequencing projects led to the development of genome annotation tools supporting the manual annotation by automated annotation workflows [150–152].

2.3.2 High-throughput sequencing

In 1988, a new sequencing approach evolved that quantified the release of pyrophosphate during DNA polymerase activity [153]. Instead of terminating DNA synthesis, this new methodology was able to constantly monitor DNA synthesis in real-time without perturbation [154,155]. In 2005, the first commercially available DNA sequencing platform entered the market taking advantage of this new sequencing protocol. This platform used an emulsion method for DNA amplification combined with a pyrosequencing protocol and triggered a new revolution; it commenced the era of the so-called NGS or *second generation sequencing* methods [71]. In the following years, several companies, *e.g.* Solexa, SOLiD and Polonator, entered the market offering commercial NGS platforms using either an emulsion PCR approach [71,156,157] or a so-called bridge amplification. The latter one building clusters of DNA fragments on a flow-cell [158] was implemented by Solexa, which was later acquired by Illumina [159].

The tremendous advances in DNA sequencing triggered by these NGS technologies revolutionized DNA research and allowed researchers to conduct experiments that were technically infeasible or unaffordable before. In 2003, the Human Genome Project published the first human genome, which took 13 years at costs of approximately US$2.7 billion [159]. Only five years later, using NGS technologies, the same has been achieved at costs of approximately US$1.5 million within five years [160]. Until the time of writing, Illumina NGS platforms made significant progress and have become the predominant NGS technology [161,162]. Until 2019, about 15,000 Illumina short-read sequencing machines were installed worldwide, which in total sequenced an astonishing amount of 150 petabases – a 50% annual increase [163]. For instance, the largest currently offered device yields an output of up to 6 terrabases and 20 billion reads in about 44 hours [164]. These steep advances in NGS techniques have led to an immense cost inflation. In 15 years, the costs of DNA sequencing using contemporary sequencing platforms have precipitously dropped from US$1 million to nowadays US$0.01 per raw megabase. For example, a 100 fold coverage of an *E. coli* genome roughly costs around US$5 [165].

Spurred by these new DNA sequencing technologies and the resulting stark rise of available DNA sequencing data, plenty of new bioinformatics methods, algorithms and software tools emerged that address the various steps required for the adequate analysis of resulting short-read data. These comprise the clipping of remaining DNA adapter sequences as well as the filtering of low-quality reads or read regions [166–169], the assembly of short DNA-sequencing reads to larger continuous sequences, *i.e.* contigs and larger scaffolds [170–175], the correction of single nucleotide or larger structural assembly errors, the filling of assembly gaps [176,177], the ordering and reorientation of contigs and scaffolds [178–183], and finally the mapping of quality-filtered short DNA-sequencing reads to reference genomes for phylogenetic analysis [184–190]. Low costs and accompanying broad accessibility of NGS technologies and bioinformatics software tools broke down barriers in terms of costs and manual efforts and thus propelled DNA based research projects. For instance, the WGS of bacterial isolates has become a standard methodology to address genomic questions. However, subsequently required bioinformatics analyses became more diverse and complex than ever due to the manifold specialized analysis steps involved. In addition, due to the short read lengths of NGS platforms, the analysis of bacterial genomes remained limited to draft assemblies, as short reads are unable to span repetitive genomic regions like ribosomal operons, insertion sequences and transposons [191–193]. Hence, complete bacterial genomes remained a demanding goal requiring manual effort and the combination of different sequencing approaches.

2.3.3 Single-molecule long-read sequencing

To overcome these limitations, a new methodology was described for the real-time sequencing of single DNA molecules in 2009 [194]. In contrast to existing NGS methods that were limited to certain numbers of sequencing cycles, this new approach used an uninterrupted template-directed DNA synthesis. This new protocol allowed the detection and constant incorporation of dye-labeled nucleotides into a growing DNA strand. A new technical platform achieved to conduct this reaction within nanostructure arrays of zeptoliter (10^{-18} ml) reaction vessels allowing the highly parallel sequencing of DNA sequences over thousands of bases without perturbation of the reaction. This new DNA sequencing technology was implemented and commercially offered by Pacific Biosciences. This platform achieved read lengths larger than 1 kilobase pairs up to several 10 kilobase pairs [194]. One drawback of this technology was the significantly lower DNA sequencing accuracy. However, due to library preparation and sequencing protocol improvements, recent devices are able to produce circular consensus sequences achieving read lengths of more than 10 kilobase pairs with nucleotide accuracies of more than 99% [195].

In 2016, Oxford Nanopore Technologies (ONT) entered the market offering an additional DNA sequencing platform. This new platform detects the sequence of single DNA molecules in real-time by measuring the current signal of ions passing a biological nanopore immobilized within a synthetic membrane along with a single stranded DNA molecule. Hereby, DNA pentamers sliding through the nanopore cause a characteristic ion current signal. This signal is measured and deciphered via bioinformatic analysis in order to identify the individual DNA base pairs [196–198]. Although the sequencing accuracy of this new technology remains considerably lower than that of Illumina short-read and Pacific Biosciences long-read sequencing technologies, lengths of resulting DNA sequencing reads eventually exceed 100 kilobase pairs. By using this new technology, in 2017 and 2018, world records for the longest sequenced DNA molecule were set achieving read lengths of 1.015 million and 1.204 million base pairs, respectively [199,200]. The latter was even argued to actually have a length of 2,272,580 base pairs that might have been incorrectly split by the ONT MinKNOW software into several subreads. One unique feature of these new devices is their small size of 10.5 cm, low weight of 87 g and low price of US$1,000 making them a truly portable and affordable sequencing platform [201]. In addition to the unprecedented read lengths, portability and accessibility, recent ONT devices produce several terabases per sequencing run [202,203]. However, challenging systematic homopolymer issues remain that cannot be fully compensated via higher sequencing coverages. However, this drawback might be extenuated or even overcome with new generations of nanopores [204]. Also, the bioinformatic analysis of the raw current signal is under active development and has made significant progress. Several different algorithms and implementations have recently been published for the initial base calling and subsequent polishing of resulting assemblies, which take advantage of hidden Markov models and deep learning techniques [205].

These new platforms again revolutionized DNA sequencing enabling tremendous advances in microbial WGS. Due to the possibility to sequence single DNA molecules in real-time, combined with the outcome of significantly longer sequencing reads, these new technologies have soon been denominated and have become referred to as *third generation* sequencing technologies [206,207]. Their great potential for microbial genome analysis led to the availability of many dedicated assemblers trying to exploit and overcome the different advantages and disadvantages, respectively [208–214]. Sequence identities of resulting assemblies can subsequently be further improved by using specialized genome polishing tools [177,211,215]. However, as none of the described sequencing technologies and related bioinformatics software tools alone is currently able to produce complete bacterial genomes with sufficiently high nucleotide identities at low costs, hybrid sequencing approaches are conducted with dedicated hybrid assemblers [216–218].

2.3.4 Assembly, annotation and characterization of bacterial whole-genome sequencing data

As explicated in the former chapters, bacterial WGS has seen tremendous progress over the last two decades. These advances from shotgun Sanger sequencing to NGS and finally third-generation real-time sequencing had large impacts on how microbes are investigated today. In only 25 years since the publication of the first complete bacterial genome sequences [133,134], high-throughput WGS has become routine and a standard methodology for many scientific applications. However, it is obvious that the mere sequencing of bacterial genomes alone is not sufficient to answer scientific questions at hand and to finally create new knowledge. Instead, this is just an initial step and many more are required to extract all the information hidden in these data. The raw data created via several technologies and provided in various data formats must be processed, analyzed and transformed into assembled and annotated genomes. This sequencing-technology-independent information can then be used for specialized *in silico* genome characterizations as well as various downstream analyses. In order to do so, multiple distinct data processing steps are required, which depend on the technology and platforms used to create the data. The following paragraphs sketch the required raw-data processing steps and possible computational genome analyses.

As a first step, potentially remaining adapter sequences are clipped from raw sequencing reads. Afterwards, sequencing reads of overall low quality are discarded and regions of low quality are trimmed. Depending on the sequencing technology and protocol, reads are filtered by length in order to discard futile too short reads and thus reduce the complexity of downstream analyses [166,167,219,220]. Afterwards, reads originating from potential vector contaminants can be detected via read mapping against dedicated databases [221]. Likewise, sequencing reads can be checked against common sources of contamination like for instance human DNA using custom databases [222,223]. Finally, DNA sequencing yields are controlled in terms of average per-base qualities, read length, remainders of adapter sequences and motif enrichments [168,224]. As a second step, these quality-filtered sequencing reads are assembled into longer contiguous sequences, called contigs. Contigs themselves might be arranged and combined into scaffolds using additional read-based information. To achieve this task, several approaches and algorithms have been described addressing short or long sequencing reads or hybrid approaches using both. For instance, to assemble short sequencing reads, overlap-layout-consensus [225] and de Bruijn graph [226,227] algorithms evolved as the predominantly used approaches. De Bruijn graph data structures proved particularly suitable to represent the overlaps of short reads by using k-mers as vertices and read paths along the k-mers as edges in the graph. Because the graph size is determined by the genome size and content of repetitive sequences, it is in principle not affected by favorably higher redundancy introduced by deeper read coverage, hence the large number of developed and available de Bruijn graph-based short-read assemblers today [170–175,228]. However, it became obvious that for the assembly of more error-prone long sequencing reads, de Bruijn graph-based approaches are not optimal and overlap-layout-consensus approaches were proven more suitable, as for example implemented in Canu [210]. One variant achieves very fast assemblies by skipping the computationally demanding consensus step implemented in Miniasm [229]. Another approach implemented in Flye [212] generalizes the idea of de Bruijn graphs to make them eligible for error-prone long reads, *i.e.* repeat graphs. Here, long reads are assembled conducting random walks through the overlap graph generating error-prone so-called disjointigs. These potentially repeated disjointigs are then collapsed into repeat representatives. The final assembly is then created by resolving these repeats via long read alignments [212]. This short exemplary list of long-read-only algorithms and assemblers is by no means complete and many more variants and other approaches exist [208,214,230,231]. A third approach is the combination of both short and long sequencing reads, which can be conducted in an either short-read-first or long-read-first manner. For the former, long reads are utilized to scaffold assembled short-read contigs and resolve loops and repeats in the assembly graph whereas for the latter, short reads are used to correct errors within long reads or resulting long-read assemblies. The short-read-first approach has been shown to provide superior results and is implemented, *e.g.* in Unicycler [216,232]. Nevertheless, none of the described assembly approaches is able to create flawless assemblies. Each DNA sequencing technology and related assembly software tools come with distinct error profiles, which has led to the development of assembly polishing tools using either short or long reads or even both. Whereas short reads are used to fix small-scale assembly errors like single nucleotide mutations and short insertions and deletions [176,233,234], long reads are used to correct medium and large-scale errors [211,235–237]. Like hybrid assembly approaches, some assembly polishing tools even implement multiple error-correction algorithms taking advantage of both data types [176,233]. Just recently, it has been shown that the combination of multiple assembly polishing tools implementing different post-assembly error correction algorithms, is able to address various error types and thus complement each other [234]. After these steps, additional quality checks are conducted in order to assess the quality of the assembled bacterial genome sequences. For this purpose, various specialized metrics and statistics like for instance the N50, have been developed as a measure for the contiguity, which is determined by the number and size of assembled contigs [238]. Another important aspect, particularly for new or rare species, is the completeness of assembled genomes. Addressing this, several phylogenomic approaches have recently evolved in order to check assembled genomes for certain single-copy orthologous genes that are common to all bacterial genomes or genomes of a distinct taxonomic lineage [239,240].

At the time of writing, the majority of assembled bacterial genomes available in the public databases resulted from short-read sequencing data. As outlined in the former chapters, these short sequencing reads cannot span the various repetitive genomic regions of bacterial genomes and thus, resulting assemblies typically remain in an unfinished status. These so-called draft genomes comprise varying numbers of contigs, typically tens to hundreds. As both order and orientation of these contigs compared to the actual biological genome sequence are determined by mere technical aspects of the implemented assembly algorithm, these contigs are required to be ordered and rearranged in a so-called scaffolding step. During this process, extrinsic genomic information from a closely related reference genome may be used to increase the synteny between the assembly and a selected reference [178,241]. To enhance this process and to expand the proportion of syntenic genomic regions between assemblies and the reference, recent scaffolding algorithms are able to utilize not just a single but multiple reference genomes [180–182]. After this step, the resulting bacterial genomes pose a common ground for many downstream analyses that are technically independent from DNA sequencing technologies.

For many of these downstream analyses, a thorough annotation of the assembled genome is required and crucial as both accuracy and comprehensiveness have strong impacts on all subsequent analysis steps. However, this process is by no means trivial. Genomic regions of interest must be either detected or predicted and then functionally described, which is denoted as regional and functional annotation, respectively. Due to the diverse genetic nature of these various genome features, an exceptionally large number of dedicated algorithms, tools and databases evolved to conduct these distinct tasks. For example, non-coding genes like tRNAs, tmRNAs, rRNAs and ncRNAs can be detected via covariance models exploiting their characteristic folding and resulting three-dimensional structures [242]. These models are collected and stored in public databases to streamline their distribution and expert curation [243]. In addition, many dedicated and more-specialized tools evolved to improve the detection, classification and functional description of tRNAs, tmRNAs [147,244,245] and rRNAs [246,247]. Besides these non-coding genes, many additional feature types can be detected via distinct tools as for example clustered regularly interspaced palindromic repeats (CRISPR) [248–253] or homology searches against specialized databases, *e.g.* origins of replication and transfer [254,255]. However, all these features combined account only for a small proportion of a bacterial genome. The majority of the bacterial genome is constituted by protein-coding genes and related coding sequences (CDS). In contrast to non-coding features, these share common characteristics, *i.e.* nucleotide triplets denoted as codons, that encode for amino acids as well as start and stop codons. These potential coding sequences between start and stop codons are called open reading frames (ORFs) that can easily be extracted from the sequence. However, due to random start and stop positions that occur on the available six translation frames resulting from three nucleotide positions in both directions of a DNA strand, ORFs tremendously outnumber actual CDS. In order to find true CDS within the vast set of all ORFs, dedicated gene prediction tools take into account additional upstream features like the presence of ribosomal binding sites and promoter sequences [146,256,257]. Then, these predicted nucleotide or amino acid sequences can be assigned to protein families and their functions can be inferred from related protein sequences. This process is denoted as functional annotation and is conducted via homology searches against databases of known sequences and subsequences that have already been described by experts. For this process, the mutual coverage and identity between two sequences are used as an approximation for homology and many dedicated algorithms and tools have been developed to solve this task as accurately and fast as possible [128,258–261]. This exemplary short list of annotation feature types only comprises higher-level and the most-important features and many more could be added, in particular regulatory regions like promoters, operators, RBS and non-coding cis-regulatory regions. Hence, the comprehensive annotation of bacterial genomes is a complex and demanding task and various centralized online services evolved to streamline the different steps that are involved in this task [150,262–264]. However, these services have become unsuitable for the timely annotation of large-scale WGS data due to the ever-increasing speed at which bacterial genomes are being sequenced today [265]. Furthermore, legal affairs and sensitive data might deem the upload to external servers inappropriate or even unacceptable. Because of these restrictions, high-throughput annotations are required to be conducted either locally on standard consumer hardware and high-performance computing (HPC) clusters or within scalable CCIs. Several command-line software tools have recently been developed to conduct this task [266–268].

Based on the genome sequence, predicted and functionally described genes and genome features, many general and more-specialized genome analyses and characterizations are feasible. One important instance thereof is the taxonomic classification of bacterial genomes. Due to their diverse nature and the fact that there is more of a continuum between bacterial genomes than clearly definable boundaries, many different *in silico* approaches evolved addressing the different taxa levels [269]. One approach is the phylogenetic analysis of the 16S rRNA gene sequence that is broadly accepted and utilized for reliable phylogenetic placements [270–272]. However, this methodology is limited in terms of resolution and thus only provides reliable placements up to the genus level. To classify genomes with higher resolutions up to the species level, whole-genome approaches like *in silico* DNA-DNA hybridization and average nucleotide identity (ANI) have become gold standards for taxonomic classifications [273–276]. For various applications, in particular outbreak detections and the surveillance of certain lineages, the species classification is often not sufficient. Hence, *in silico* MLST analyses are conducted to detect sub-species lineages and assigned sequence types are compared and shared world wide [85,277]. Besides the taxonomic classification and genome typing, annotated genomes and genes pose an invaluable starting ground for countless downstream analyses. With regard to modern medical microbiology and epidemiology, the detection, annotation and surveillance of AMR genes and virulence factors are of particular interest and utmost importance to the field and thus, led to the development of many software tools [278–283] and databases [282,284–288] that are available today to achieve these tasks. Another example for specialized gene-based analyses is the detection and annotation of biosynthetic gene clusters [289,290]. Of note, gene-based approaches are often complemented by sequence based methodologies in order to detect and annotate genome features that are otherwise hard to identify as for example insertion sequences [291–293].

In conclusion, this tremendous progress in the field of bacterial WGS gave rise to a plethora of highly-specialized bioinformatics algorithms, software tools and databases. Today, researchers must choose from multiple alternatives in order to conduct the various explicated tasks of data processing, related downstream analysis and genome-based computational characterizations. In addition to detailed per-genome characterizations, the broad availability of both tools and WGS data meanwhile fosters the standardized analysis of entire cohorts of multiple closely related genomes. However, this constant progress, which cannot be considered to be attenuating [206,294], implies new challenges that must be addressed. Besides data processing steps and genome-based analyses, additional mere technical issues emerge, as for example rising requirements for data storage, management and transfer [40]. Furthermore, the various data processing and analysis steps need to be combined and integrated into automated and reproducible workflows that are executed in a scalable manner on different computing resources and infrastructures [295].

2.4 Antibiotic-resistant bacteria – a global threat and challenge for public health

> "Then there is the danger that
> the ignorant man may easily underdose himself
> and by exposing his microbes to
> non-lethal quantities of the drug,
> make them resistant."
>
> Sir Alexander Fleming
> Nobel Prize acceptance speech, 1945

The serendipitous discovery of penicillin in 1928 by Alexander Fleming was a milestone in medical history [296]. In the 1940s, its clear antibacterial effects and non-human toxicity led to large-scale production and mass treatments. Meanwhile, Selman Waksman achieved to turn this incidental discovery into a standardized screening procedure for molecules with antibiotic effects and introduced the technical term *antibiotic* [297]. Both discoveries have been awarded with well-deserved Nobel Prizes as these findings triggered the biggest medical revolution since the discovery of vaccines. Accompanied with the steep scientific progress made during the mid-1900th century, it was then thought that bacterial diseases would be easily controlled and the threat of many infectious diseases would finally come to an end. However, already in his Nobel Prize acceptance speech Alexander Fleming stated that "there is the danger that the ignorant man may easily underdose himself and by exposing his microbes to non-lethal quantities of the drug, make them resistant". Even though, not underdosing but quite the opposite happened, he was unfortunately proven right in the very same year by the first detection of penicillin resistant bacteria. In the "golden era" of antibiotics around the 1960s, most today-known classes of antibiotics have been discovered. Only a few classes of antibiotic drugs have been found thereafter, *e.g.* daptomycin – the last one in 1986 [298].

Nevertheless, these fairly easy discoveries of antibiotic drugs created an uncritical and wasteful use without decent considerations of the potential consequences, which later on materialized [299]. Between 2010 and 2015, the aggregated consumption of antibiotics in 76 countries increased by 65% from 21.1 to 34.8 billion defined daily doses – a common metric to measure antibiotic consumption [300]. This large-scale human consumption is comprehensively described and well-understood in many different setups and there is compelling evidence that it is a primary driver for the emergence of antibiotic resistances [301–305]. Too often, the effectiveness of these necessary drugs is threatened by the unnecessary prescription by physicians uncertain of diagnoses and consumers lacking better knowledge or unaware of the problem [306]. Comparatively, these disastrous developments are even dwarfed by the mass usage of antibiotics in livestock farming. The global consumption of antimicrobial agents was recently estimated at a level of about 63,000 tons. Even worse, the global usage is projected to increase by 67% to approximately 105,000 tons by the year 2030 [307]. The soaring global demand for meat driven by the rising global population and desire of low- and middle-income countries to catch up with the often critical lifestyle of the western countries, led to meat production growths since 2000 of 68%, 64%, and 40% in Africa, Asia, and South America, respectively [308]. This demand fosters antimicrobial usage in order to increase livestock productivity, which equals a large share of 73% of the entire global consumption of antimicrobials [309]. The considerable and nearly constant contact of animals in livestock farms with antimicrobial drugs establishes favorable conditions for the selection of antibiotic-resistant bacteria and might provide important reservoirs for antibiotic resistance genes [307]. This immense evolutionary-active selection pressure caused by the mass prescription and consumption of hundreds and thousands of tons of antibiotics has put the world at the dawn of a post-antibiotic era that would pose an equally severe and tragic medical regress of almost an entire century [306].

Unfortunately, these developments are not the only concerns for public health. During the last 50 years, at least 26 emerging bacterial infectious agents have been identified. A key driver for this increased exposure of humans to bacterial pathogens are major changes in human lifestyle as well as the constant rush into and settling of previously uninhabited rural nature for both industrial and leisure reasons. Our natural environment is a sheer infinite prokaryotic reservoir and many of those might play a pathogenic role once transmitted from their ecological niche to humans [55,310,311]. A large proportion of bacterial diseases originally derive from animals, livestock or wildlife, and thus are considered as zoonoses. Since 1940, 60% of 335 emerging infectious diseases events were zoonoses, 54% are attributable to bacteria [312]. This combination of emergence of new bacterial pathogens on the one hand and the emergence of novel antibiotic resistance genes on the other hand constitute a global issue for public health worldwide.

For instance, in Europe in 2007, nearly 400,000 infections have been estimated and about 25,000 deaths have been attributed to only six antibiotic-resistant bacterial pathogens resulting in approximately 2.5 million extra hospital days [313]. The implied costs of hospitalization and loss of productivity summed up to total costs of 1.5 billion €. This estimated burden further increased to nearly 670,000 infections and about 33,000 attributed deaths in 2015 [32]. Likewise, a recent study of the CDC from 2019 estimates 2.8 million infections with antibiotic-resistant bacteria in the USA, leading to approximately 35,000 deaths, annually [314].

This worldwide surge of multi-resistant bacteria has led to the realization that without the implementation of effective countermeasures, in 2050 up to 10 million people could die annually due to infections with antibiotic-resistant bacteria [30]. To address these huge medical threats, many new drugs have been introduced, which have been evolved via modifications of existing antibiotic targets. However, this repertoire of effective drugs found in the golden era of antibiotics has run short and global pharmaceutical companies became reluctant to invest in their antibiotic drug pipelines. This lack of economic investment and research exacerbated the precarious situation of available antibiotic drugs [315]. Because of these developments, it has become obvious to scientists, the health-care community and policymakers, that new antibiotic targets and approaches are urgently needed [316]. In order to guide research, discovery and development of new antibiotics, the World Health Organization (WHO) published a global priority list of antibiotic-resistant bacteria posing the most-severe threats to public health, *e.g.* *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacteriaceae* [317]. Some of the most dangerous bacterial pathogens have become famous as the so-called antibiotic-resistant ESKAPE pathogens: *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species [318].

In 2019, over 400 scientific projects from more than 300 institutions worldwide actively investigated new antibiotic targets and drugs. A large proportion hereof follows entirely new approaches enabled by recent findings, which in turn are only possible because of the large amount of deeply characterized bacterial genomes as well as novel epidemiological knowledge gained by large-scale genome analyses [319]. Although there are considerable global efforts towards the discoveries of new antibiotic targets and approaches, again it has become obvious and common sense that a shift of the common mindset is required in order to fight back antibiotic resistances. Accelerating the required pace of the global community, the WHO urgently advocated for a global action plan in 2015 bringing together scientists and policymakers [320]. The explicated set of countermeasures and described paths for a more sustainable economy in regards to the usage of antibiotics in public health, veterinary and agriculture has recently been acquainted with the term *ONE health* [321,322].

One effective and necessary countermeasure is the early containment of antimicrobial-resistant pathogens stopping the spread of emerging or highly prevalent AMR genes. In order to do so, a deep understanding of the underlying antibiotic resistance mechanisms (Figure 3b) as well as their epidemiology at different scales is required to forestall unnecessary prescriptions. The immense increase of recently sequenced antibiotic-resistant bacterial genomes has impressively shed light on the diverse nature of antibiotic resistance mechanisms and their genetic determinants (Figure 3c), which can be grouped into innate and acquired resistances. Innate resistances originate from spontaneous genetic mutations modifying cellular targets (Figure 3a) of antibiotic drugs attenuating or stopping susceptibility to these. This comprises for instance, point mutations in the 16S ribosomal RNA gene conferring resistance to tetracycline derivatives [282] and alterations in the regulatory machinery leading to increased or decreased transcription rates of resistance targets [323]. These innate resistances primarily disseminate via vertical gene transfer. Besides, acquired resistances denote the active or passive incorporation of new genes that cause antibiotic resistances into the genome via horizontal gene transfer, such as conjugation, transduction and transformation [324]. Main mechanisms hereof are mobile genetic elements, *e.g.* plasmids, transposons, integrons, conjugative elements and bacteriophages [325–327]. These mobile genetic elements are key drivers for the spread and evolution of antibiotic resistance genes. Except for bacteriophages, mobile genetic elements fall in two categories: those that can move from one bacterial cell to another and those which can move from one genetic location to another within a cell [327]. This often results in complex genetic landscapes harboring resistance genes within nested mobile elements, which therefore are able to move from one system to another [328].

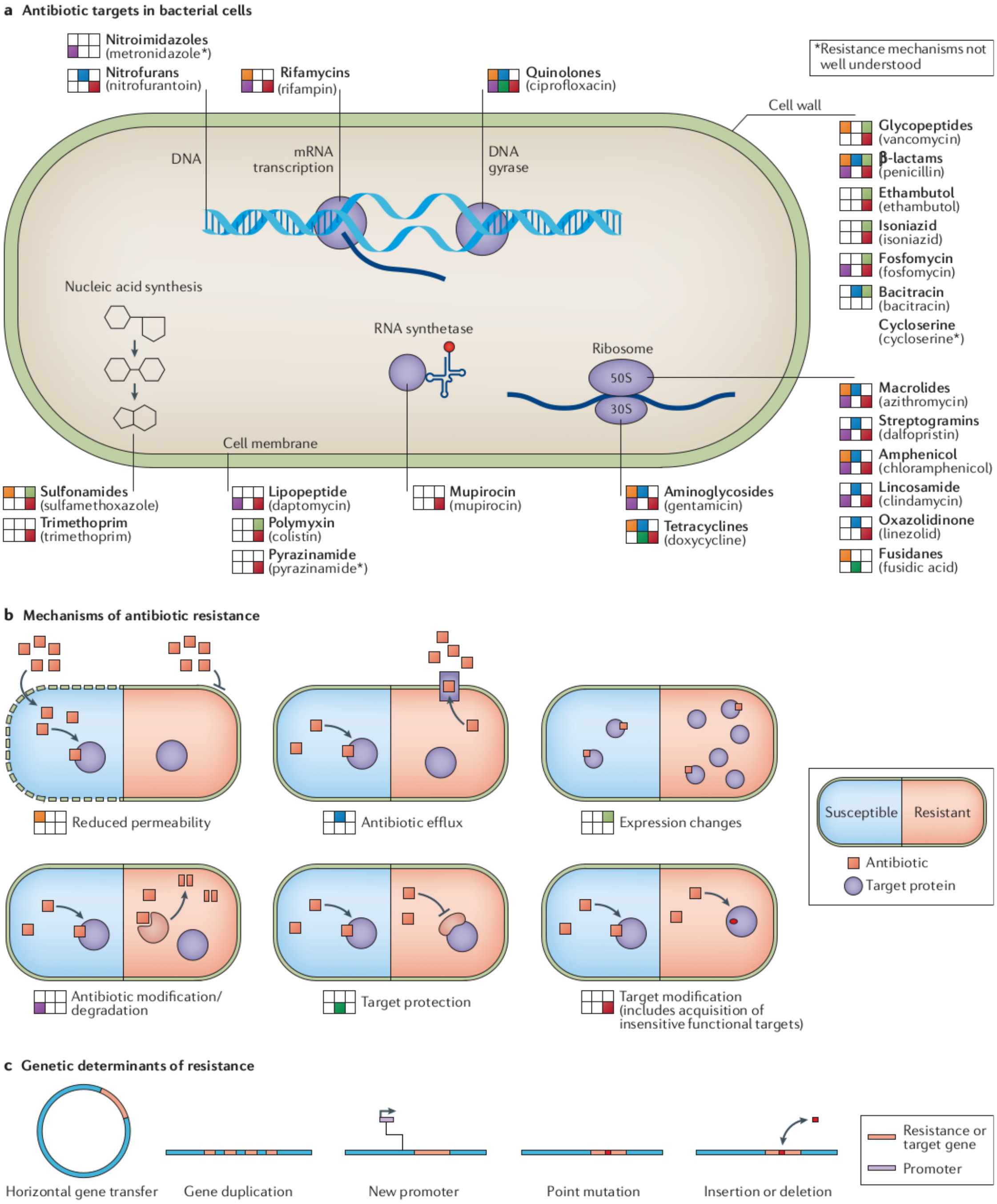

a Antibiotic targets in bacterial cells
Nitroimidazoles (metronidazole*)
Nitrofurans (nitrofurantoin)
Rifamycins (rifampin)
Quinolones (ciprofloxacin)
*Resistance mechanisms not well understood
Cell wall
DNA
mRNA transcription
DNA gyrase
Nucleic acid synthesis
RNA synthetase
Ribosome
50S
30S
Cell membrane
Glycopeptides (vancomycin)
β-lactams (penicillin)
Ethambutol (ethambutol)
Isoniazid (isoniazid)
Fosfomycin (fosfomycin)
Bacitracin (bacitracin)
Cycloserine (cycloserine*)
Macrolides (azithromycin)
Streptogramins (dalfopristin)
Amphenicol (chloramphenicol)
Lincosamide (clindamycin)
Oxazolidinone (linezolid)
Fusidanes (fusidic acid)
Sulfonamides (sulfamethoxazole)
Trimethoprim (trimethoprim)
Lipopeptide (daptomycin)
Polymyxin (colistin)
Pyrazinamide (pyrazinamide*)
Mupirocin (mupirocin)
Aminoglycosides (gentamicin)
Tetracyclines (doxycycline)
b Mechanisms of antibiotic resistance
Reduced permeability
Antibiotic efflux
Expression changes
Susceptible
Resistant
Antibiotic
Target protein
Antibiotic modification/ degradation
Target protection
Target modification (includes acquisition of insensitive functional targets)
c Genetic determinants of resistance
Horizontal gene transfer
Gene duplication
New promoter
Point mutation
Insertion or deletion
Resistance or target gene
Promoter

Figure 3: Exemplary depiction of targets, mechanisms and genetic determinants of antibiotic resistances.

a) Groups of antibiotic drugs act on various molecular targets within bacterial cells. b) Antibiotic resistance is implemented by numerous molecular and genetic mechanisms. c) Genomic alterations and genetic mutations as determinants of antibiotic resistances. Reprinted with permission from Nature Reviews Genetics [323], Copyright © 2019, Springer Nature.

This large number of different genetic determinants and acquisition mechanisms make the surveillance and epidemiological tracing of resistance genes a delicate task. For example, the large and diverse group of β-lactamases is a prominent example for their complex heterogeneity. Today, β-lactamases are categorized by different classification systems, *e.g.* Amble and Bush-Jacoby-Medeiros using either protein sequence homologies or phenotypic profiles, respectively [329,330]. According to the Amble classification, β-lactamases are classified into four groups: A, C, D representing serine β-lactamases and group B representing metallo-β-lactamases [330]. Each group comprises many subgroups with variants denoted by different naming schemes, as for example TEM named after Temoneira – the first patient from which samples harboring these alleles were collected, CTX and OXA named by their primary antibiotic drug targets cefotaxime and oxacillin, KPC named by the species *Klebsiella pneumoniae* and NDM named after New Delhi – the location of its first detection [331]. The rigor description, categorization and typing of antibiotic resistance gene alleles is a crucial task for the surveillance of emerging genes as well as the epidemiological tracing of their dissemination. In 2017, more than 1,800 variants [331] of β-lactamase protein sequences have been described of which many occur globally. One example is the NDM metallo-ß-lactamase group, which has been reported for the first time in 2008 in a patient isolate in New Delhi conferring resistance to a variety of penicillins and cephalosporins. Only three years later, many of its derivatives are reported worldwide [332]. Bad enough, but the rise and spread of ß-lactamases, which include extended spectrum ß-lactamases, is only one example. Due to these omnipresent resistances, other and mostly newer antibiotics are reserved as last-resort drugs, *e.g.* colistin. However, just recently a plasmid-encoded resistance gene called *mcr-1,* which was initially found in China [26], has now been detected all over the world, *e.g.* Laos, Thailand, Nigeria, Europe [25] and Germany [27].

The described disseminations of antibiotic resistance genes are prominent examples of the global efforts to understand the emergence and spread of these genes based on modern WGS technologies. Increased bacterial sequencing projects driven by further cost reductions and streamlined bioinformatic analysis pipelines are expected to contribute to the rapid inhibition of further disseminations and hopefully real-time outbreak detections on a global scale, soon. Though, WGS approaches are not limited to the detection of antibiotic resistance genes alone. Another application of utmost importance is the *in silico* prediction of AST. However, in 2017 the European Committee on Antimicrobial Susceptibility Testing (EUCAST) reviewed the current development status of WGS for bacterial antimicrobial susceptibility testing (AST) and came to the conclusion that there is a lack of evidence that WGS could be used for AST in clinical settings today. Amongst many issues, more quality controls, performance standards and common comparative measures are necessary [333]. However, despite these open issues, DNA sequencing technologies and related protocols in clinical environments have come a long and astonishing way considering what can nowadays be achieved by these methodologies [334]. Once these issues are overcome, whole-genome based *in silico* prediction of antibiotic susceptibility might be a fast and cost-efficient alternative [335].

The steep progress in molecular biology and genetics led to an understanding of the underlying molecular targets (Figure 3a) of antibiotic resistances, *e.g.* the gyrase involved in DNA replication, polymerase involved in mRNA transcription, ribosomes involved in the protein translation as well as the cell membrane and cell wall. These targets are involved in the many known mechanisms as for instance a reduced cell permeability, antibiotic efflux, expression changes, target protections and enzymatic modifications and degradations of antibiotics (Figure 3b). The various underlying genetic determinants (Figure 3c) for all of these molecular mechanisms of a phenotypic resistance can be grouped into two aforementioned distinct classes, innate and acquired resistances. The latter are predicted via the identification of a certain gene that is known to infer a resistance via homology searches against resistance gene databases and the application of decent quality thresholds, *e.g.* the mutual sequence coverage and identity of query and subject sequences. Over the last decade, more than 15 public databases of AMR genes emerged and evolved, not including many additional species-specific databases [34,323]. However, the mere detection of an antibiotic resistance gene alone often does not provide sufficient information for an accurate phenotype prediction as many biological processes, as described above, can have an important effect, either in *cis* or *trans* location. These processes form the group of innate resistances, which are notoriously hard to predict, as many different molecular targets and often complex mechanisms are required to be taken into account. Besides the described technical hurdles, which hopefully and most-certainly will be overcome soon, these very complex genetic determinants pose a severe challenge and open field for modern bioinformatics and currently hamper the precise AST *in silico*.

Still, as more and more deeply sequenced and phenotypically characterized genomes of bacterial pathogens are available, potentially combined with transcriptomic, proteomic and metabolomic data, the numerous molecular and genetic interplays and dependencies can be comprehensively depicted in order to foster a deeper understanding of complex antibiotic resistance mechanisms. One example of such deep analysis, which achieves very high genetic resolutions, is a recent study investigating the genetic evolution of antibiotic determinants in different bacterial in-patient isolates by a combination of genomic and transcriptomic data [336]. Another promising but also demanding approach is the exploitation of modern machine-learning techniques for the analysis of large amounts of genotype and phenotype datasets in so-called genome wide association studies (GWAS). Recently, many studies have been described that address genomic AST by machine-learning approaches [337]. For example, Chen *et al.* investigated genetic variants of 28 targeted genomic regions from more than 3,600 phenotypically described *Mycobacterium tuberculosis* genomes. Next to favorable achievements in the prediction of resistance phenotypes, by doing so they could also identify previously uncharacterized mutations as important predictors for certain resistance types potentially pointing to new antibiotic targets and mechanisms [338]. Especially, the identification and interpretation of machine-learned genetic determinants of innate resistances is gaining more attention, as these are notoriously hard to identify via classical approaches. For instance, a recent study published a biochemically interpretable machine-learning classifier for microbial GWAS putting the available genetic data into the context of biochemical pathways [339]. A further example of machine-learning approaches is the single nucleotide polymorphism (SNP) based prediction of antibiotic susceptibility. High-quality data given, recent studies could show that decent predictions are possible [36,338,340] potentially even predicting distinct susceptibility levels, *i.e.* the minimal inhibitory concentration for certain drugs [341]. However, the requirement for large amounts of high-quality data poses a notable hindrance to these promising approaches. In order to unleash their full potential, vast numbers of sequenced genomes along with high-quality phenotypic characterizations are required and thus will further drive and increase the rate of bacterial WGS [323]. Furthermore, this high sequencing rate combined with phenotypic characterizations must be continued in order to forestall genomic data and actual phenotypes drifting apart and to keep databases up to date regarding new antibiotic determinants. In addition, further progress in DNA sequencing technologies in terms of costs and throughput will drive the sequencing of even more isolates of given samples. This in turn would provide the foundation for deeper analyses of intra-host evolutions and intra-population variations on the smallest scales down to the level of single cell sequencing.

2.5 *In silico* detection of bacterial plasmids

Plasmids are genetic vehicles and constitute an important mechanism for both vertical and horizontal gene transfer in bacteria and thus play a vital role in the spread of genes within and between bacterial populations [342–345]. Genes encoded on plasmids comprise a large genetic repertoire often featuring non-essential metabolic, resistance and virulence capabilities that provide an evolutionary advantage in certain environments [346–348]. A prominent element in this group with large medical and epidemiological implications are genes conferring resistance to antibiotic drugs. Many acquired antibiotic resistance genes are actively or passively mediated via plasmids between bacterial organisms and additionally exchanged between plasmids and the chromosome via transposons and integrons [326]. These often complex and versatile genetic landscapes foster the spread of resistance genes, which has been traced and comprehensively described in the literature based on DNA sequencing techniques [349]. A prominent example thereof is the traced global spread of the plasmid-encoded *mcr-1* gene inducing resistance against colistin – a last-resort antibiotic drug. This gene has been initially found in *Enterobacteriaceae* isolated from human and animal samples collected in China [26]. Later, the *mcr-1* gene was found widely spread over the whole world, *e.g.* Laos, Thailand, Nigeria and Europe [25]. In 2017, it has additionally been described to be detected in Germany [27], as well. Therefore, the plasmid-mediated spread of antibiotic resistance genes is an issue of increasing severity. It is well known that plasmids are able to break species boundaries and thus spread widely, for instance via wildlife both taxonomically and geographically [350].

Hence, the automated screening of bacterial genome assemblies for the presence of plasmids is a necessary and important task and a powerful tool for plasmid-based epidemiology. It is broadly known and accepted that DNA-based *in silico* approaches for the identification and characterization of plasmids provide profound advantages in terms of sensitivity and specificity over classical molecular methodologies. However, depending on the used DNA sequencing platform, this requires several necessary bioinformatics tasks that introduce new challenges. Due to the complex genetic landscape of bacterial genomes and in particular plasmids, short-read WGS approaches regularly fail to recover complete chromosome and plasmid sequences. This is caused by repetitive regions like rRNA operons, insertion sequences and transposons, which are known to notoriously hamper finished short-read assemblies [191–193]. Nowadays, many of these issues can be addressed and often solved by long-read sequencing technologies providing sequencing reads that are long enough to span these repetitive regions. However, new issues arise from these technologies in turn, as the advantage of longer DNA-sequencing reads comes at the cost of comparatively lower read quality in terms of sequence identity. For some sequencing technologies, this is even exacerbated by higher rates of systematic sequencing errors that cannot be fully compensated by higher sequencing depths. Because of these issues, long-read-only assemblies are still unsuitable for many standard epidemiological *in silico* analysis, *e.g.* multi-locus sequence typing, resistance allele typing and transmission studies. Additionally, long-read sequencing libraries are often filtered *in silico* for longer sequencing reads that help closing short-read assemblies. Unfortunately, discarding the typically large number of shorter long reads in turn often results in losing small plasmids. It could also be shown that small plasmids tend to be underrepresented in some long-read DNA libraries that are optimized for larger DNA fragment sizes and thus further exacerbate these issues [351]. For these reasons, high-accuracy short-read data is still required for such applications [352]. To address and finally overcome these issues, hybrid sequencing approaches combining short and long-read technologies emerged that triggered the development of dedicated assembly tools, as for instance a recently enhanced SPAdes version and Unicycler, which improves and complements the SPAdes assembly workflow [216,218]. However, as long-read sequencing platforms are still notably less cost-efficient compared to short-read sequencing platforms, most large-scale WGS projects still rely on the latter. Furthermore, long-read technologies emerged in only recent years and thus, public DNA repositories still provide significantly more short-read WGS data for mere historic reasons aside from cost effects. Hence, a necessary first step in many *in silico* plasmid analysis workflows is the detection and extraction of plasmid-borne contigs from short-read draft assemblies posing a binary classification problem: either a contig originates from the chromosome or a plasmid.

This classification problem is a bioinformatic challenge and has resulted in multiple new approaches and many implementations of many software tools. These address either the identification or even the entire reconstruction of plasmid sequences within bacterial whole-genome short-read draft assemblies. They can be divided into three categories (Table 1). The first comprises tools searching for known genes and related subsequences in highly specialized databases. PlasmidFinder, for example, detects DNA subsequences of genes necessary for the flawless plasmid replication machinery, which are known as incompatibility groups [353]. MOB-suite seeks to identify conserved relaxase protein sequences, which are necessary for the mobilization of plasmids, from a highly curated and dedicated database [254]. A second large group comprises tools analyzing k-mers and varying k-mer frequencies. PlaScope [354] and PlasmidSeeker [355] conduct lookups against pre-built databases whereas cBar [356], PlasFlow [357], mlPlasmids [358] and PlasClass [359] take advantage of more elaborated machine-learning approaches in order to exploit subtle frequency differences and complex non-linearities hidden in the data. The third group comprises a heterogeneous set of tools analyzing assembly graphs. Short-read assemblies almost never result in closed genomes but complex graph structures representing potential paths through connected contigs. Many tools take advantage of this additional information, which is provided by contemporary assemblers as assembly graphs. PlasmidSPAdes [360] and Recycler [361] exploit k-mer coverage variations between contigs. Recycler [361], PLACNETw [362] and gplas [363] take into account additional information from paired-end reads that bridge disjoint contigs and try to find circular paths through the assembly.

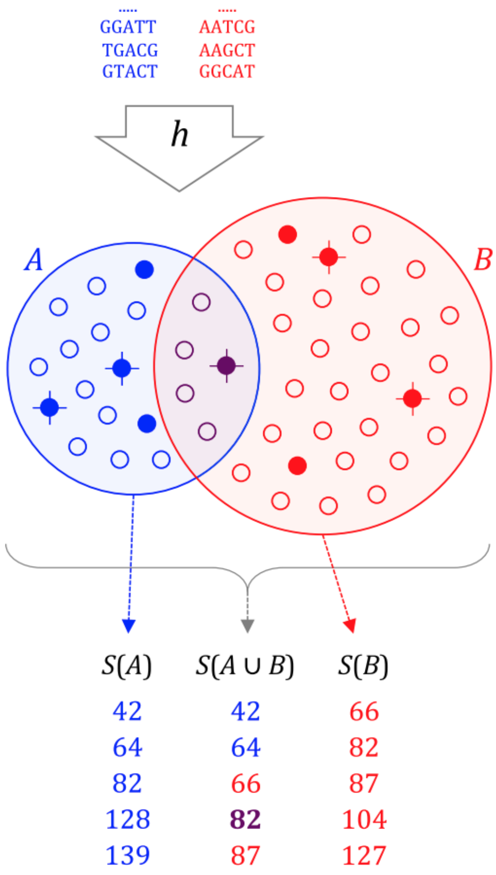
Table 1: Approaches, methodologies and tools for the detection of plasmid-borne contigs within bacterial draft assemblies.

Category	**I**	**II**	**III**
Input type	Assembled genomes		Sequencing reads and/or assembly graphs
Approach	Detection of conserved genes and sequence probes from curated databases	K-mer frequency analysis	Assembly graph analysis
Methodology	Homology searches	Statistics, machine learning	Statistics, heuristics
Tools	PlasmidFinder, MOB-suite	PlaScope, cBar, PlasFlow, mlPlasmid*, PlasClass	PlasmidSPAdes, Recycler, PLACNETw*, gplas
Disadvantage	Low sensitivity	Targeted databases	Dependency on sequencing technology

* Non-automated interactive workflow

All these approaches come with distinct advantages and shortcomings making the optimal tool selection a difficult task. None of the described approaches achieves a combination of reasonably high sensitivity and specificity, but are rather biased towards one or the other. Furthermore, many follow targeted approaches addressing particular taxa, favor certain plasmid sizes or ranges of sequencing coverage [364–366]. These limitations complicate the selection of tools and methodologies and make them inadequate for the integration into untargeted, fully automated and sequencing technology-independent analysis workflows for the large-scale analysis of bacterial WGS data.

2.6 Applications for reference genomes and optimal selections thereof



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Figure 4: Overview of the MinHash approach to approximate the Jaccard index.

A pair of input sequence sets is decomposed into two sets of k-mers. These sets of k-mers are transformed into sets of hashes using a hash function. Hashes are sorted according to their numeric values and for each set, the *m* lowest hashes are selected as a sketch, *i.e.* a representative set of k-mers. Finally, the fraction of common hashes between the two sketches and the *m* lowest hashes of both sketches is used as an approximation of the Jaccard index. Reprinted with permission from Genome Biology [374], Copyright © 2016, Springer Nature.

Since the introduction of the very first nucleotide databases in the 1980s [118–120], the number of publicly available genomes is constantly rising. For example, between 1982 and today (2021), the number of genomic sequences stored in GenBank increased from 606 to 219,055,207. Of note, the number of unfinished sequences resulting from whole-genome short-read sequencing projects increased from 172,768 in 2002 to 1,517,995,689 today [367]. Therefore, the average yearly growth rate of WGS sequences of ~60% considerably outpaced the average yearly growth rate of complete sequences of ~40%. This tremendously increasing number of available genome sequences, especially microbial including bacterial, led to the realization that curated, high-quality non-redundant collections are necessary in order to manage, maintain and represent the large and diverse taxonomic range of available microbial sequences. Over the last decade, tremendous efforts went into the design, setup and maintenance of such reference sequence repositories regarding both assemblies and annotations [264,368,369]. These representative genome sequences as well as certain genomes, which have been analyzed and described particularly well *in vivo*, *in vitro* as well as *in silico*, are generally denoted as reference genomes and used in many types of downstream analyses.

For instance, a required task for many downstream analyses is the mapping of sequencing reads onto a common genetic region in order to analyze similarities and differences on various levels, *e.g.* single nucleotide polymorphisms, insertions, deletions as well as structural variations [185,187,370]. High-quality reference genomes provide these common genetic regions along with additional genomic context via annotations. An important application of mapped sequencing reads is the single nucleotide polymorphism detection for the subsequent calculation of phylogenetic trees to analyze pathogenic clonal outbreaks [371]. Another example is the reference guided assembly of short sequencing reads [372]. In contrast to *de novo* assemblies, reference-guided assemblers take into account extrinsic genomic information of sufficiently related reference genomes. A further very important processing step after the assembly is the ordering and rearrangement of contigs within draft assemblies. As assemblers have no or only constraint information regarding the actual order and orientation of assembled contigs, so-called scaffolding software tools are used to map these contigs onto closely related reference genomes to reconstruct their most likely order and orientation [179–182].

Hence, the selection of suitable reference genomes has become an important and critical pre-analysis task as this choice has large impacts on downstream analyses [373]. In order to compare and rank available reference genomes regarding their distance to a certain query genome, several *in silico* methodologies emerged, *e.g.* comparison of tetranucleotide frequencies, Genome BLAST Distance Phylogeny, ANI and k-mer-based Jaccard indices [273–276,374]. Inspired by well-established *in vitro* hybridization of DNA fragments, the alignment of DNA subsequences of a certain genome against another genome with subsequent computations of average identities and conserved values has been shown to robustly represent the relatedness between both [273]. Hence, the computation of ANI and conserved DNA values has been implemented in several online and offline tools [274,375,376]. However, applied on larger numbers of bacterial genomes, the large computational effort caused by pairwise alignments of the DNA subsequences is a crucial drawback of this methodology. A faster approach is the alignment-free comparison of k-mer sets via the Jaccard index, *i.e.* the fraction of common k-mers. However, comparing millions or even billions of k-mers is still a demanding task. But, this process can be accelerated and requirements for computational resources can be reduced via the computation of MinHashes using approximations of the Jaccard index between two genomes (Figure 4). For each genome, all canonical k-mers are hashed, sorted and reduced to a subset of a given size, which is denoted as a sketch. For each pair of sketches, an approximation of the Jaccard index is calculated and provided as a measure of genome relatedness that was shown to correlate well with more precise alignment-based ANI values [374,377]. However, this correlation depends on k-mer lengths, sketch sizes as well as the genomic distances between genomes. Here, the reduction of runtimes, which is achieved by reducing the amount of compared genomic content, comes at the cost of reduced resolution for closely related genomes. In order to mitigate these drawbacks, a k-mer based alignment-free implementation of ANI computations was recently published [276]. However, this implementation in turn is not applicable to genomes that are too distantly related to each other.

Hence, for each analysis a decision must be made to choose between methodologies taking into account available computing resources, runtime requirements and genome distances. A further common drawback of all available implementations is the necessity to compile a database of reference genomes and a lack of integrated taxonomic information and metadata. Of note, thoroughly calculated and sufficiently low whole-genome distances to well described reference genomes, are also an eligible methodology for the taxonomic classification of bacterial genomes.

2.7 Recent IT developments and challenges for microbial bioinformatics in the 2020s

The game-changing developments in DNA sequencing revolutionized the way how genomic data is created and how many bacterial genomes are routinely sequenced every day. These large amounts of genomic data represent a scientific treasure trove providing huge potentials and possibilities. Large-scale comparisons of hundreds and thousands of bacterial genomes can be used to investigate within-host diversity and evolution [378,379], to delineate and reconstruct local outbreaks [380–382], to describe global population structures and to answer epidemiological questions [74,383]. However, these massively growing numbers of sequenced genomes also introduce new challenges. All this DNA sequencing data must be properly processed and analyzed on its own and effectively compared against each other [384]. New standards and standard operating procedures for bioinformatics data processing and analyses are necessary in order to effectively compare samples analyzed using different sequencing technologies by different laboratories worldwide. Accelerated by the tremendous cost inflation of DNA sequencing, it has recently been estimated that the yearly acquisition of DNA sequencing raw data could rise to a worldwide level of one zettabyte in 2025 [40]. A well-known phenomenon and symbol of technological development is Moore’s law. It states that the number of transistors fitting on an integrated circuit board, a rough equivalent for CPU power, is increasing exponentially, with a doubling time of approximately 18 months [385]. Of note, this ventured prediction held true for more than 35 years. A similar prediction exists for the storage capacity of hard drives; Kryder's law predicts the hard drive storage capacities to double every 12 months [386]. For a long time, this constant technological progress of computational capabilities and capacities had easily kept pace with the requirements of DNA sequencing and related bioinformatics. However, since the advent of the NGS technologies in the middle of the first decade of this century, the technological progress in the field of high-throughput DNA sequencing vastly outgrows Moore’s and Kryder’s laws [295]. Between 2008 and 2016, the capacity of DNA sequencing platforms doubled, on average, every seven months [40]. Large genome projects, *e.g.* the 100,000 Genomes Project [387], the Human Microbiome Project [388] and the Earth Microbiome Project [389], tremendously increase the size of public DNA data repositories. As no climax of this trend can be anticipated in the foreseeable future, it might threaten the centralized dogma of contemporary global genome and sequencing raw data repositories. The mere amount of data will soon make it infeasible to upload all unprocessed raw data into centralized repositories and therefore, increasing the demand for local raw data processing on the one hand and scalable, distributed and nearby computing infrastructures for large-scale analysis of potentially pre-processed data on the other hand [295]. Meanwhile, cloud computing has evolved for the last two decades as a new paradigm for such compute infrastructures. According to the US National Institute of Standards and Technology (NIST), a formal definition of cloud computing, is “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources … that can be rapidly provisioned and released with minimal management effort or service provider interaction” [390]. Starting in the early 2000s, computational resources were initially offered on demand. Operated and maintained within large data centers, physical machines were shared among different users via virtual machines, which were globally accessible via the Internet. By sharing the physical computational backbone between many different users, synergistic effects, *e.g.* ceased overprovisioning for peak loads and sharing spare resources, reduce costs and the necessary know-how to build and maintain larger IT infrastructures [391]. This new way how IT resources are provided and used gained momentum over the last decade, when large technology companies entered the market. Beside computing resources on demand, more and more software tools were provided as a centralized service known as *software as a service*. This shift in the IT ecosystem also partly transforms the way how DNA sequencing data as well as genomic data is processed and analyzed. This process in turn will require bioinformatics software tools to be developed in a high-throughput-savvy and scalable manner, *i.e.* they are either executable on local computers or deployable to scalable CCIs.

Driven by many large-scale academic sequencing projects as well as applied research projects in modern microbial biotechnology [392–394], the tremendously increased amount of available data opened new scientific questions and constantly required new algorithms and bioinformatics approaches. This contributed to the genesis of a plethora of open-source bioinformatics software tools and databases. At the time of writing, the online registry bio.tools, compiled as part of the European Infrastructure for Biological Information (ELIXIR), includes 17,276 entries from over 2,462 contributors [395]. Among those, 6,186 were annotated with DNA sequencing and genetics-related terms. Often, these highly specialized tools need to be combined and integrated into more comprehensive analysis workflows. As each of these tools has its own set of requirements and software dependencies, the provisioning of these workflows has become a non-trivial task.

Hence, in order to isolate software applications from their environment, *e.g.* the operating system, installed software libraries and available third-party software tools, software containers have recently evolved as a lightweight new mechanism of isolation and thus also portability. In contrast to virtual machines, which require their own full stack of operating system, libraries and software applications, containers run within the kernel of the operating system. Thus, they require less resources and provide better performance [396,397]. Taking advantage of these containerization techniques, developers and researchers are able to package and execute software tools combined with all dependencies in a lightweight and portable manner, across a wide range of computing platforms [398–400]. Furthermore, these container images can be uploaded to and distributed via public repositories, as for instance, Docker Hub. Today, many distinct containerization systems exist. Among others, the most successful and widely used are Docker [41] and Podman [42].

In addition to these issues regarding the isolation, packaging and distribution of single tools, contemporary computational analysis workflows also have to deal with a broad set of technical runtime issues. Robust implementations of analysis workflows are expected to provide reproducible results on different machines over multiple iterations. Furthermore, the analysis of large datasets requires workflows to be executed in a fault-tolerant manner. Corrupted data parts or failed executions of the analysis of some parts of the data must not lead to failures and crashes of the entire workflow. In these cases, workflow implementations are expected to properly handle such failures and to further proceed with the computation. Otherwise, problematic parts of the data or failed parts of the underlying computing infrastructure might hinder the completion of the entire workflow. The latter is of particular importance on fault-tolerant infrastructures like CCIs. An additional aspect, and maybe the most important, is scalability. Modern analysis workflows need to be applicable on a broad range of data sizes. During the developmental stage, small test cases need to return quick results, while in a production stage, real analyses might scale to very large amounts of data. The latter often requires to either scale vertically or horizontally by distributing the computational workload to compute clusters of different types and varying sizes. It goes without saying that the support of multiple HPC cluster systems [401–403] and state-of-the-art cloud computing frameworks [404–406] facilitates increased portability of implemented workflows and therefore, also their applicability. Hence, in order to decouple the scientific methodological development of analysis workflows from the outlined issues of the mere technical execution, many dedicated workflow engines for bioinformatics use cases have been developed and evolved. Among many others, two of the most recent, feature rich and widely used are SnakeMake [407] and NextFlow [408].

The steep progress in DNA sequencing technologies and the resulting data inflation require bioinformaticians to keep pace with the complex demands of modern software tool development in bioinformatics on the one hand and the heterogeneous technical solutions on the other hand. However, this novel layer of added technical complexity will be worthwhile as by mastering scalability and portability challenges, a democratization of bioinformatics and computational biology is taking place and thus removes historical obstacles implied by unavailable or too complex computational resources [398]. Due to public CCIs, as for instance the German de.NBI cloud [409], small research groups and even single researchers are able to conduct data analyses at almost any scale. After the democratization of research data via the broad availability in public data repositories, finally a democratization of data analysis is taking place, too.

2.8 Scientific gaps in microbial bioinformatics and aim of this thesis

The former chapters provide a brief historical introduction to the huge scientific and technological progress in the field of microbiology, especially medical microbiology, boosted by the revolutionary developments of high-throughput DNA sequencing technologies and accompanying bioinformatics methodologies, software tools and databases. Today, DNA based bioinformatic analyses have become essential and powerful research tools for the field of microbial genomics addressing various scales from single genomes to population structures and ecosystems. In addition, DNA based *in silico* analyses have become invaluable tools for public health tasks like for example the surveillance of bacterial pathogens, AMR monitoring and outbreak detections [35,36,410]. Many of these applications either benefit from or even require the processing and higher-level characterization of multiple genomes in order to determine genetic commonalities or differences between genomes. These developments are important factors that increase the demand for large cohorts of collectively analyzed bacterial genomes. Meanwhile, over the last decades advances in DNA sequencing led to massive cost reductions and a tremendous increase in sequencing throughput (Figure 5). Hence, large-scale bacterial genome sequencing has reached unprecedented levels and has become a standard methodology and a routine task in laboratories worldwide. Today, public databases comprise hundreds of thousands of bacterial genomes. For example, in 2018, the European Nucleotide Archive (ENA) stored more than 660,000 genomes. Of note, only 20 pathogenic species are accountable for more than 90% of these genomes, which underpins the huge importance of these technologies and data for medical microbiology [411].

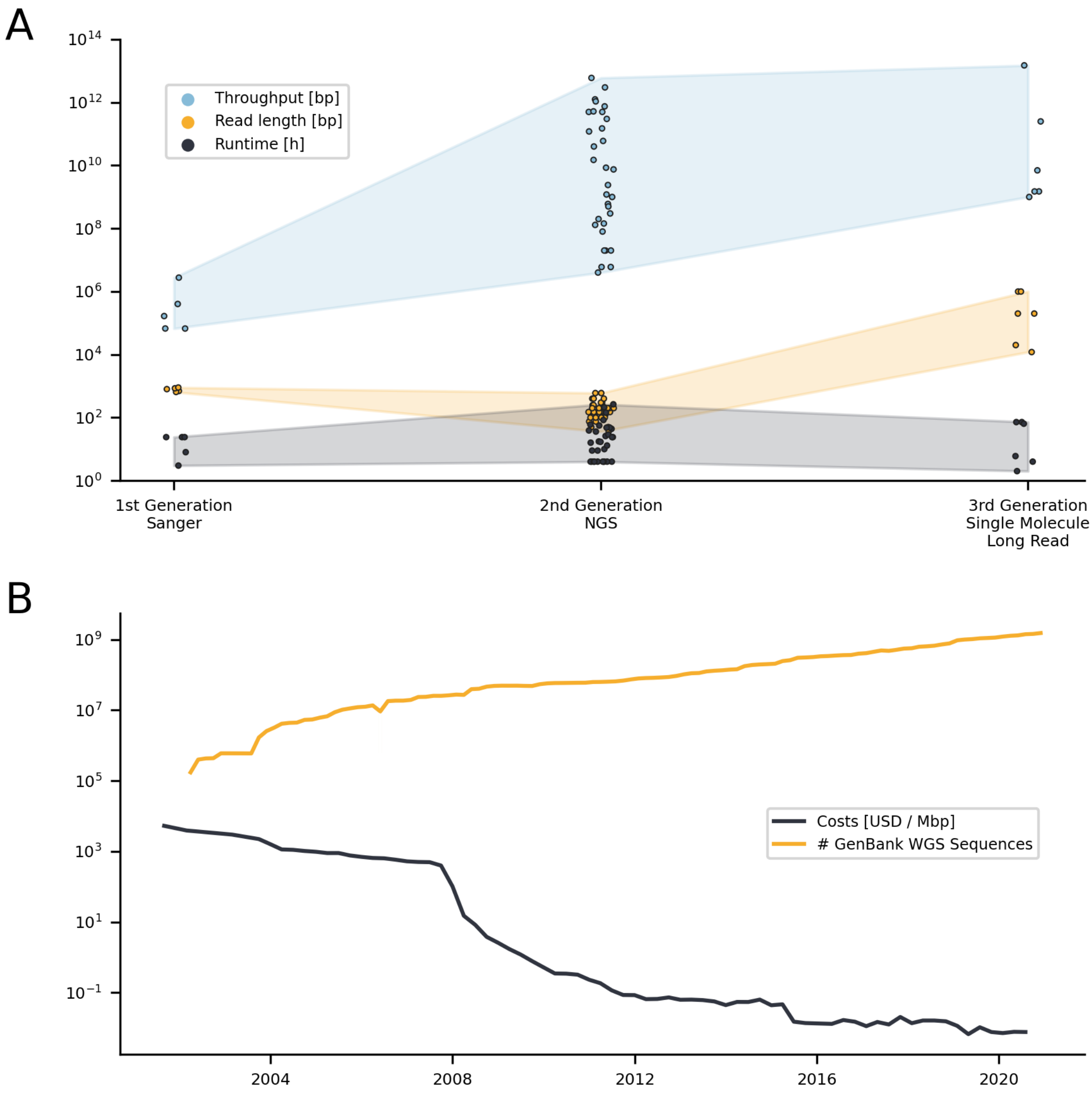


Figure 5: Progress in DNA sequencing.

Advances in DNA sequencing technologies by time and technological revolutions. A) Throughput, average read length and runtimes of 42 modes of commercially available DNA sequencing devices grouped by the underlying DNA sequencing technology, *i.e.* Sanger sequencing, next-generation short-read sequencing and single-molecule long-read sequencing. Distinct data points are depicted as circles. General trends are highlighted as coloured bands according to each characteristic [412–415]. B) The temporal course of DNA sequencing costs in USD per million base pairs and the number of whole-genome sequences stored in the NCBI GenBank database [416,417].

As the sequencing of bacterial genomes is evidently not a limiting factor anymore, it has become obvious that the effective and efficient analysis of all this data is becoming a new bottleneck. The sheer amount of available and newly generated data has made the manual analysis a tedious and time-consuming process, which thus is becoming more and more infeasible. Furthermore, the repetitive manual execution of similar tasks is a common source of errors and potentially insufficient standardizations are an important aspect regarding reproducibility and comparability. Hence, the comprehensive analysis of this data has become a very complex task. Appropriate analysis workflows comprise many steps of which each constitutes a distinct niche in bioinformatics by means of methodology as well as software implementation. Today, researchers can, but also have to, select bioinformatics tools from countless choices comprising thousands of specialized software tools [395]. This is further exacerbated by the fact that most software tools provide a large set of options and parameters to fine-tune their performance and behavior and to optimize the outcome of an analysis. These often require highly specialized domain knowledge and significant experience. Even worse, in order to conduct many of these tasks, bioinformatics software tools must be executed in combination with specialized databases. For example, at the time of writing, there are at least 15 publicly available AMR gene databases [323]. This poses an increasing problem for researchers. Results of WGS data analyses become more and more incomparable by the usage of different workflows composed of different software tools using different sets of options and parameters potentially in combination with different databases leaving out that many of these are provided in regularly updated releases.

Accordingly, there is a rising demand for standardization in bacterial WGS data analysis for the sake of reproducibility of conducted analysis and comparability of results. Raw sequencing data from varying DNA sequencing platforms must be processed and analyzed and subsequent results are to be aggregated and prepared to create human readable reports facilitating the rapid and comprehensive understanding of the results (Figure 6). To overcome these issues, automated and centralized analysis platforms recently emerged, *e.g.* Bacterial Analysis Pipeline [418], Patric [287] and Galaxy [419]. These platforms provide researchers with access to complex analysis workflows that are executed on centralized IT infrastructures via convenient web user interfaces thus hiding most of the implied scientific and technical complexity. However, centralized platforms cannot constantly keep pace with the steep increase of generated data resulting from decentralized sequencing sites worldwide. In addition, the transfer of large amounts of raw data is physically limited by public network capacities. Furthermore, sensitive data is often not eligible for the analysis on third party infrastructures due to legal restrictions.

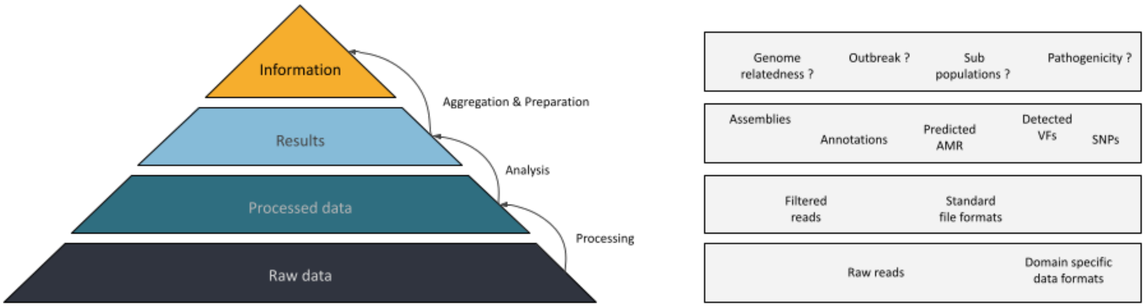


Figure 6: Transformation from raw data to information

Depiction of the transformation from raw data into information with examples in regards to microbial bioinformatics. Large-scale raw data resulting from different DNA sequencing platforms must be processed and analyzed in various ways to create new results. To gain new information, these diverse results must be aggregated, prepared and finally presented in comprehensible manners.

Consequently, there is a need for automated and comprehensive but also portable analysis pipelines that can be executed either locally on standard consumer hardware and HPC clusters or CCIs in a scalable manner. However, there is a lack of bioinformatics software pipelines fulfilling all these requirements. Hence, it was the aim of this thesis to address this gap by developing a new bioinformatic analysis pipeline for the automated, comprehensive and scalable analysis of small to large cohorts of bacterial WGS data from different DNA sequencing platforms. The following chapters briefly describe the requirements of the involved tasks that resulted from this objective. The first task was the design and implementation of the analysis pipeline. The second and third tasks resulted from the design of its fully automated workflow and addressed the automated detection and characterization of plasmids and the rapid but thorough determination of suitable reference genomes.

2.8.1 Standardized high-throughput analysis of whole-genome sequencing data of bacterial cohorts

Driven by the inflation of sequenced bacterial genomes in many different settings, *e.g.* academia and public health, the first task of the described objective is the development of a new bioinformatics software tool for the analysis of bacterial WGS data fulfilling the following requirements:

- Design of a fully automated, standardized, reproducible and comprehensive data processing and analysis workflow
- Support for WGS data from all major contemporary DNA sequencing platforms, *i.e.* Illumina, Pacific Biosciences and Oxford Nanopore Technologies
- Execution of a comprehensive set of per-isolate genome characterizations
- Implementation of comparative and phylogenetic analyses
- Vertical and horizontal scalability on local hardware, HPC clusters and CCIs to keep pace with rising amounts of data
- Portability and user-friendly installation routines on standard consumer hardware.
- Extensibility via a modular framework design
- Compilation of human-readable, user-friendly and interactive hypertext markup language (HTML) reports aggregating, preparing and visualizing intermediate and final results

2.8.2 Automated and taxonomy-independent detection of plasmid-borne contigs from short-read draft assemblies

Due to their important role in the horizontal transfer of resistance genes, a crucial aspect of bacterial WGS data analysis is the detection and characterization of plasmids, which has been addressed by a large number of dedicated bioinformatics software tools that have recently evolved [354–358,360–363,420]. However, despite the previously described heterogeneity of plasmid detection methodologies and software tools, none of these provide all properties that are required for the seamless integration into a contemporary and automated WGS data analysis workflow described in task I:

- A non-interactive and thus fully automated workflow
- An underlying classification approach that is purely based on assembled draft genomes providing a common workflow entry point for the support of various DNA sequencing platforms
- Untargeted and taxonomy-independent workflow and database
- High detection accuracy achieving balanced sensitivity and specificity

Compliance with all these requirements would make such a methodology applicable to WGS data from a large range of bacterial taxa supporting different sequencing platforms and thus would allow the automated separation of plasmid-borne contigs from the chromosome for focused and more detailed downstream analyses. Hence, it was the second task to develop a new methodology fulfilling the outlined requirements and to implement this new approach as an automated bioinformatic software tool for high-throughput applications.

2.8.3 Accurate but rapid determination of suitable reference genomes

The deep characterization of bacterial isolates on a nucleotide level is an important task to understand phenotypic differences between strains caused by single nucleotide variants (SNVs) and SNPs, which are detected against closely related common reference genomes. Moreover, SNPs that have been called against a common reference genome pose a well-accepted method for the calculation of phylogenetic trees with utmost precision down to each single nucleotide. Here, reference genomes act as a genetic template masking non-common genetic information. Hence, the selection of suitable closely related reference genomes is an essential task with large implications for the results of SNP based analysis. Another application requiring even more than a single reference genome is the ordering, rearrangement and scaffolding of assembled contigs. Modern scaffolding software tools are able to utilize combinations of different genomic landscapes from several reference genomes for the rearrangement and optimal placing of contigs [180–182]. Both examples are essential parts of the comprehensive workflow described in the objective of this thesis and task one. However, contemporary software tools [274–276] for the assessment of potential reference genomes do not fulfill all of the required following properties:

- A locally executable command line implementation
- A fully automated workflow
- Short runtime while still achieving high-quality results
- Integrated databases comprising public high-quality genomes

Hence, it was the third task to develop a rapid, accurate and integrated bioinformatic software solution for the fully automated lookup of suitable reference genomes.

3 Thesis contributions

This thesis comprises three peer-reviewed publications, which are presented and briefly summarized in the following subchapters.

- **ASA³P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates.**
  Oliver Schwengers, Andreas Hoek, Moritz Fritzenwanker, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).
  PLoS Computational Biology, DOI: 10.1371/journal.pcbi.1007134

- **Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein-sequence-based replicon distribution scores.**
  Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).
  Microbial Genomics, DOI: 10.1099/mgen.0.000398

- **ReferenceSeeker: rapid determination of appropriate reference genomes.**
  Oliver Schwengers, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).
  Journal of Open Source Software, DOI: 10.21105/joss.01994

3.1 ASA³P

ASA³P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates.

Oliver Schwengers, Andreas Hoek, Moritz Fritzenwanker, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).

PLoS Computational Biology, DOI: 10.1371/journal.pcbi.1007134

This publication presents and describes ASA³P, a new bioinformatic software tool for the comprehensive analysis of WGS data from bacterial isolates. ASA³P implements a state-of-the-art fully automated analysis workflow comprising the quality control and assembly of raw reads, scaffolding and annotation of resulting assemblies and the thorough characterization of bacterial isolates. The latter comprises taxonomic classifications and subtyping, the detection of AMR genes and virulence factors, and the detection of SNPs. These per-isolate analyses are complemented by comparative analysis, *i.e.* the computation of core and pan genomes and phylogenetic trees. Of note, ASA³P supports all contemporary major sequencing platforms, *i.e.* Illumina short-read sequencing as well as Pacific Biosciences and Oxford Nanopore Technologies long-read sequencing. ASA³P is publicly available and provided as two distinct software distributions. Small to medium cohorts can be locally analyzed via Docker-based Linux containers, whereas large to massive groups of up to thousands of isolates can be analyzed with a highly scalable cloud computing version, which is able to fully exploit the flexibility and scalability of modern CCIs. Finally, results are provided in standard bioinformatics file formats and gathered information is presented via user-friendly reports comprising interactive visualizations. It has been demonstrated that the software smoothly scales from small to very large datasets comprising more than 1,000 genomes, and it has been successfully applied in various data analysis projects, which are described in chapter 4.1.2.

3.2 Platon

Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein-sequence-based replicon distribution scores.

Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain,
Trinad Chakraborty & Alexander Goesmann (2020).

Microbial Genomics, DOI: 10.1099/mgen.0.000398

Platon is a new bioinformatic command line tool for the fully automated detection, characterization and extraction of plasmid-borne contigs from bacterial draft assemblies. Via large-scale homology searches of publicly available closed chromosome and plasmid sequences, it could be shown that a large proportion of bacterial proteins is unequally encoded within the different replicon types. A new statistical score termed replicon distribution score (RDS) reflecting this bias for each marker protein sequence (MPS) is defined and introduced as a new methodology to approach this problem. Via RDS, Platon is able to exploit this natural distribution bias for the determination of the origin of contigs, which is further enhanced by heuristics taking into account higher-level contig characterizations as for example: circularization tests, detection of incompatibility groups, mobilization and conjugative genes, detection of origin of transfer sequences, detection of ribosomal genes, and homology searches against plasmid reference databases. Final results are provided in standardized human and machine-readable file formats for user-friendly examination as well as automated downstream analysis. Platon was shown to achieve higher accuracies and more robust classifications in taxonomy-independent benchmarks and better or equal performance on targeted benchmarks than existing tools.

3.3 ReferenceSeeker

ReferenceSeeker: rapid determination of appropriate reference genomes.

Oliver Schwengers, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).

Journal of Open Source Software, DOI: 10.21105/joss.01994

This publication describes the new bioinformatic software tool ReferenceSeeker that allows researchers to query large microbial genome databases for closely related high-quality reference genomes in a rapid and integrated manner. The software implements a two-step search process combining the rapid lookup of candidate reference genomes from integrated local databases via k-mer fingerprints with detailed computation of the well-established average nucleotide identity and conserved DNA values. Even more detailed comparisons can be conducted via the computation of bidirectional ANI values. Due to this generic approach, the software supports a broad range of microbial taxa. Furthermore, the software allows the creation of customized databases incorporating non-public genomes. Pre-compiled databases comprising NCBI RefSeq genomes are publicly available via open data repositories for bacteria, archaea, fungi, protozoa and viruses.

4 Results and discussion

This thesis provides three scientific contributions to the field of microbial bioinformatics addressing the highly relevant explicated issues: the analysis of large-scale bacterial WGS data, the automated prediction of plasmid-borne contigs from draft assemblies and the rapid determination of suitable reference genomes from custom and public databases for high-throughput applications (Figure 7).

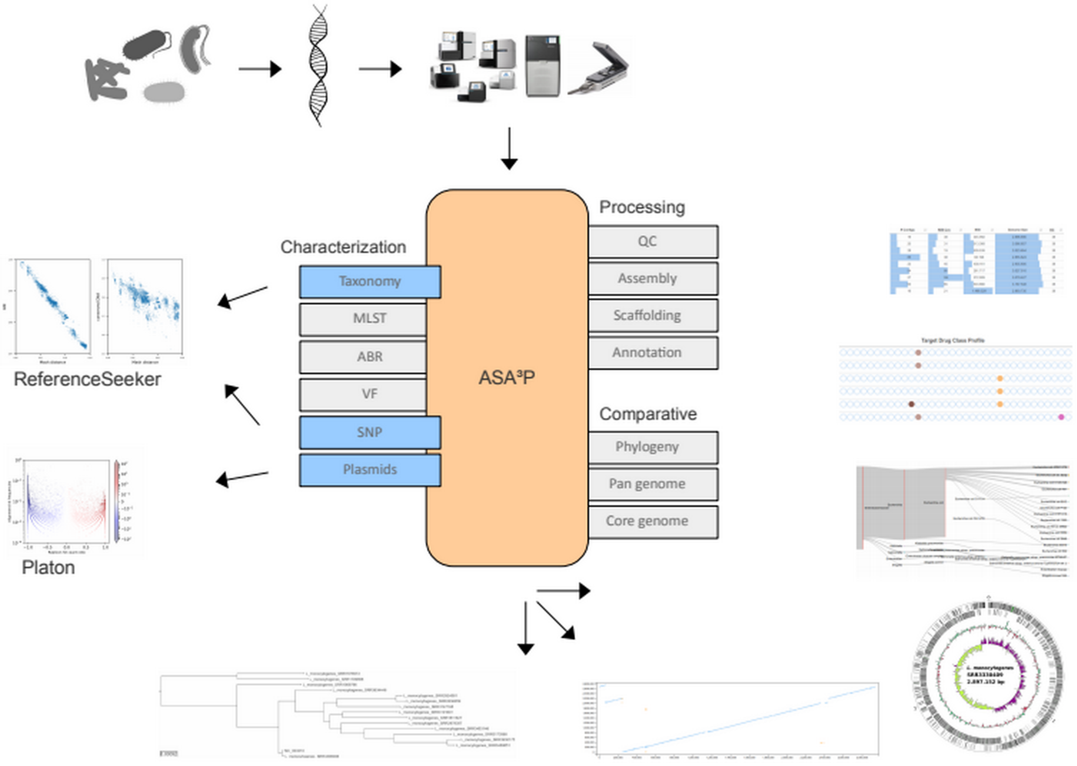


Figure 7: A comprehensive and fully automated analysis workflow for bacterial WGS data.

Depicted is the comprehensive analysis workflow developed in this thesis starting after upstream wet laboratory steps. Raw sequencing reads are processed by ASA³P resulting in assembled and annotated bacterial genomes that are further analyzed and characterized. Results are provided as human-readable interactive reports. Open bioinformatic challenges that emerged from the design of this workflow have been addressed by new methodologies and approaches implemented in the software tools Platon and ReferenceSeeker.

4.1 Comprehensive, scalable and fully automated high-throughput analysis of whole-genome sequencing data from bacterial isolates with ASA³P

4.1.1 Features and comparison with contemporary software tools

Triggered by the constantly increasing computational demands caused by the immense developments in high-throughput DNA sequencing, a comprehensive analysis workflow has been designed and implemented in a scalable, locally executable and portable manner resulting in the software tool ASA³P. It provides a true one-stop solution lightening the burden of repetitive bioinformatics analysis tasks. By design, the pipeline provides no adjustable parameters to the user and by doing so, enforces the standardization of the analysis and in turn the reproducibility as well as comparability of results. These are generated in standard bioinformatics file formats and stored in a well-defined and predictable file structure suitable for subsequent custom analysis. In addition, the software generates user-friendly interactive reports as standard HTML documents, which can easily be compressed and sent to colleagues and research partners and viewed with common web browsers. Hereby, accessible and comprehensible higher-level insights to the underlying data are provided to users without the need for sophisticated bioinformatics or Linux command line skills.

Moreover, support for the analysis of raw sequencing reads from Illumina, Pacific Biosciences and ONT sequencing platforms increases the overall usability of the software. The conducted analyses as well as generated reports cover a comprehensive set of contemporary bacterial genome characterizations as for example the taxonomic classification and MLST subtyping, the prediction of antibiotic resistance genes and detection of virulence factors, as well as SNP-based comparisons to reference genomes. The software was implemented following a modular design and therefore, it can easily be expanded with further analysis modules.

Depending on available IT capacities and the number of sequenced genomes, users can choose from a locally executable Docker-based version and an OpenStack-based cloud computing version. Hence, by providing these distinct software distributions, ASA³P is able to scale from the analysis of tiny bacterial cohorts executed on regular desktop machines to the analysis of thousands of bacterial genomes within large distributed HPC clusters or CCIs. Hence, ASA³P enabled researchers to keep pace with the demands of contemporary bacterial WGS data analysis.

Recently, several bioinformatics software tools and pipelines for the automated offline analysis of bacterial WGS data have evolved following different approaches to meet distinct requirements. Although all available tools provide a comparable set of core analysis features, each tool has different properties in terms of supported data types, scalability and flexibility and thereby addresses distinct requirements. For example, BacPipe [421], Tormes [422], Nullarbor [423], rMAP [424] and ProkEvo [425] only support short-read sequencing data. Bactopia [426], for instance, supports the hybrid assembly of short-read and long-read sequencing data. However, it does not support the assembly of long sequencing reads, only. In contrast to ASA³P and Tormes, the user interfaces of Bactopia, BacPipe and Nullarbor provide adjustable parameters in order to adapt the underlying workflows and analysis steps. Bactopia and BacPipe generate results in bioinformatics file formats only whereas Tormes, Nullarbor and ASA³P generate human readable reports as markdown, static and interactive HTML files, respectively. All software tools but BacPipe provide vertical scalability and Bactopia and ASA³P additionally provide horizontal scalability supporting HPC clusters and CCIs. Furthermore, all software pipelines but Tormes provide portable Linux container images.

In conclusion, albeit each software provides its unique set of properties, at the time of writing, ASA³P is the only available open-source bioinformatic software tool providing a fully automated and comprehensive analysis workflow, support for both short and long read data assembled in either separate or hybrid mode, vertical and horizontal scalability on HPC and CCIs, and the generation of comprehensive and user-friendly reports in a portable manner for the offline analysis of bacterial WGS data. Thus, it enables researchers to take advantage of scalable IT resources and a diverse set of robust and proven bioinformatics software tools dedicated to the various tasks involved in the process. Hence, even more bacterial genomes and larger cohorts thereof can be analyzed, characterized and compared allowing to keep up with DNA sequencing technologies and future demands.

4.1.2 Examples of application

Long before its publication, ASA^{3}P has already been widely used as an inhouse analysis pipeline at the department for Bioinformatics and Systems Biology and the Institute of Medical Microbiology within several scientific projects in the context of the German Center for Infection Research (DZIF). Over the course of the recent years, more than 8,400 sequenced genomes of bacterial pathogens have been analyzed covering a broad taxonomic range comprising more than 50 distinct genera [427,428]. Table 2 lists the 15 most frequently analyzed genera comprising many severe pathogens, among these all of the so-called ESKAPE species [429]. From the analysis of these small and large-scale data, various scientific findings have been published in the field of medical microbiology based on the cohort analyses of antibiotic-resistant bacteria conducted with ASA^{3}P.

Table 2: Number of pathogenic bacterial isolates analyzed with ASA³P within the various subprojects of the DZIF grouped at the genus taxon. Listed are the 15 most frequently analyzed genera. [427].

Genus	Number of analyzed isolates
Escherichia	3,597
Klebsiella	872
Enterococcus	840
Serratia	764
Enterobacter	501
Listeria	224
Pseudomonas	196
Citrobacter	191
Acinetobacter	184
Proteus	96
Staphylococcus	68
Actinobacillus	62
Streptococcus	42
Hafnia	25
Moraxella	18

For instance, *E. coli* isolates from Nigerian and Ghanaian poultry farms as well as from hospitals in Germany and Switzerland have been analyzed and characterized regarding prevalent taxonomic sublineages and AMR genes [430–432]. Similarly, ASA³P has been used to study clinically relevant bacteria collected from German surface waters [28] – among these *E. coli*, *A. baumannii*, *K. pneumoniae*, *C. freundii*, *E. cloacae* and *P. aeruginosa*. Also, ASA³P has been used to analyze and delineate isolates of *L. monocytogenes* [433]. In all these studies, the entire *in silico* data processing, analysis and characterization was conducted using ASA³P to provide the relevant information within the distinct studies' contexts, *e.g.* the clonal subtyping using multi-locus sequence typing as well as the detection of antibiotic resistance genes.

In addition to these cohort analyses, in several studies ASA³P facilitated the deep characterization and comparison of distinct strains of a certain species of interest. For instance, WGS data from isolates of *Bordetella pseudohinzii,* a new atypical species causing respiratory infections like whooping cough, has been analyzed in order to deeply characterize the sequenced genomes via detected SNPs against public reference genomes. Furthermore, intermediate ASA³P results from pre-processed WGS data were used to be further analyzed by more specialized downstream analysis tools [434]. Likewise, ASA³P has also been used to deeply characterize different capsule mutants of *Streptococcus pyogenes* on a per-SNP basis against related wildtypes [435]. A further example is the analysis of *Enterobacter bugandensis* – a pathogen that causes severe infections of neonates, which has been isolated in Germany for the first time [436]. WGS data of sequenced samples was processed, assembled in a hybrid approach, annotated and characterized using ASA³P. In total, ASA³P has been cited 28 times due to Dimensions.ai, at the time of writing.

4.1.3 Ongoing developments and pending challenges

ASA³P conducts a comprehensive and state-of-the-art workflow to process and analyze bacterial WGS data. However, several important aspects of bacterial WGS analysis remained untouched, which are subject of current and future developments. This certainly pertains to the detection and analysis of mobile genetic elements like prophages and transposons as they play a vital role in the dissemination of antibiotic resistance genes and potentially have large impacts on phenotypes [326,328,345,437,438]. If both scientifically and technically feasible, these mobile genetic elements should be well characterized and compared against each other.

For example, the detection and characterization of plasmids is of very high relevance for the analysis of bacterial genomes. In order to automatically detect plasmids within genome assemblies while simultaneously supporting different sequencing technologies, the underlying plasmid detection methodology is required to be able to work on the assembled genome sequences alone. From the many bioinformatics software tools for the automated plasmid sequence detection publicly available today, only a tiny fraction works solely on draft assemblies [353,354,356,357]. These, however, do not satisfy the manifold requirements of this automated, taxonomy-independent, integrated, multi-sequencing-platform analysis workflow, as they either require species specific databases [354] or predict plasmid sequences strongly biased towards sensitivity or specificity [353,356,364]. To overcome this issue, the development and implementation of a new plasmid-born contig detection methodology fulfilling the explicated requirements became an interesting and challenging objective during the design of the ASA³P workflow. As a result, a new methodology for the robust detection and characterization of plasmid-borne sequences within bacterial draft genomes became part of this thesis and is described in the following chapter 4.2.

In addition, further issues became obvious while designing and implementing the workflow of ASA³P. In order to calculate reference genome SNP-based phylogenetic trees, a single particular reference genome is required and must be provided by the user. However, it might not be clear which reference genome fits best the data at hand. This issue is exacerbated by the sheer overwhelming number of publicly available reference genomes, which is constantly rising. Hence, it would be beneficial to the results of the conducted analyses as well as to the overall usability, if appropriate reference genomes could be tested, assessed and finally chosen to be included in the analysis in an automated manner if no reference genomes are provided by the user. Furthermore, as modern scaffolding software tools are able to take advantage of multiple reference genomes, the automated selection of larger numbers of closely related reference genomes might further improve scaffolding results of short-read draft assemblies [179,180,182]. Hence, the automated determination of suitable reference genomes was one objective of this thesis and is addressed in more detail in chapter 4.3. The described hierarchical approach has already been re-implemented in Groovy and integrated into ASA³P as a replacement for the k-mer based taxonomic classification via Kraken [439]. Compared to Kraken, this new taxonomic classification approach is based on well-established ANI and conserved DNA species boundary thresholds computed against a tightly integrated compilation of reference genomes. As a positive side effect, this resulted in a database storage size reduction from 142 GB to 29 GB and thus, significantly reduced overall resource requirements.

Another example for potential future enhancements is the annotation of assembled bacterial genomes. At the time of writing, ASA³P takes advantage of the widely accepted software tool Prokka due to its streamlined command line interface and short runtimes. The latter is achieved by using hierarchical annotation databases exploiting annotated reference genomes, which are aggregated to the genus level. However, this information must be provided via a distinct parameter, which therefore, is required to be provided by ASA³P users. A taxonomy-independent annotation software tool would allow to relax or even remove this requirement. Furthermore, the annotation workflow conducted by Prokka exhibits certain limitations, *e.g.* the detection and annotation of small CDS, a large proportion of CDS annotated as hypothetical protein especially in rare species and the proper detection and annotation of CDS spanning artificial replicon edges.

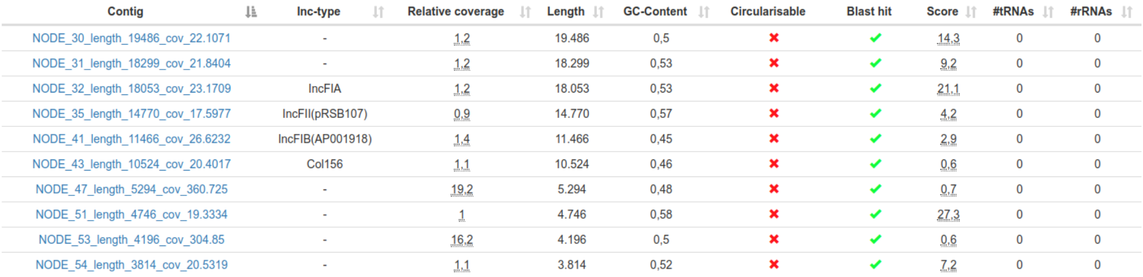
4.2 Robust detection and characterization of plasmid-borne contigs from bacterial draft assemblies with Platon

Plasmids are vital vehicles for bacterial genes. As a key mechanism of horizontal gene transfer, plasmids play an essential role in the dissemination of antibiotic and metal resistance genes. In order to monitor and understand the role of plasmids within single genomes as well as their dynamics within bacterial populations, nowadays, the detection and characterization of plasmid sequences via the bioinformatic analysis of bacterial WGS data has become an essential tool. However, due to the often complex and nested composition of different mobile genetic elements, short-read assemblies are hardly ever complete but fragmented, comprising multiple contigs. These fragmented sequences make the *in silico* detection and characterization of plasmids a difficult task [328,364]. Recent bioinformatics software tools for the detection of plasmid-borne contigs from WGS assemblies do not fulfill the complete set of requirements for fully automated, non-interactive, scalable and taxonomy-independent analyses without the necessity to choose between either sensitivity or specificity.

Hence, a novel methodology was developed achieving the outlined requirements for the seamless integration into ASA^{3}P's workflow. As a new approach to this problem, differential distributions of protein-coding gene families among chromosomes and plasmids were investigated by large-scale analysis [440]. It could be shown that a considerable proportion of these protein sequences is significantly unequally distributed among replicons. This inherent natural bias is used to classify contigs and to predict a replicon's origin. The conducted benchmarks show that this methodology achieves a superior classification performance compared to both taxonomy-independent [356,357] as well as targeted approaches [353,354]. Furthermore, it is applicable without any adaptations or customizations within both scenarios as its classification was proven to be the most sensitive whilst still achieving a specificity close to the most specific approaches [353]. It was implemented as a stand-alone bioinformatic software tool providing contemporary plasmid characterizations providing useful additional information, *e.g.* sequence circularity and incompatibility factors.

Moreover, due to a fully automated database creation workflow, Platon's mandatory taxonomy-independent database can regularly be updated without manual efforts. Thus, the increasing amount of sequences that are stored in public genome repositories can be utilized to constantly keep RDS values of MPS up to date and thus, forestall databases from becoming outdated. Furthermore, the incorporation of more complete chromosome and plasmid sequences will further improve the predictive power of MPS and their RDS values. Indeed, a benchmark on 1,765,157 simulated contig sequences alike those conducted in the publication showed that Platon's classification performance was further increased by a recent database update incorporating replicon sequences of RefSeq release 202. The highest contig classification accuracy defining the RDS conservative threshold (CT) that was achieved, could be further increased compared to the published software release v1.2.0.

4.2.1 Integration into ASA³P



Contig	Inc-type	Relative coverage	Length	GC-Content	Circularisable	Blast hit	Score	#tRNAs	#rRNAs
NODE_30_length_19486_cov_22.1071	-	1,2	19.486	0,5	✗	✓	14,3	0	0
NODE_31_length_18299_cov_21.8404	-	1,2	18.299	0,53	✗	✓	9,2	0	0
NODE_32_length_18053_cov_23.1709	IncFIA	1,2	18.053	0,53	✗	✓	21,1	0	0
NODE_35_length_14770_cov_17.5977	IncFII(pRSB107)	0,9	14.770	0,57	✗	✓	4,2	0	0
NODE_41_length_11466_cov_26.6232	IncFIB(AP001918)	1,4	11.466	0,45	✗	✓	2,9	0	0
NODE_43_length_10524_cov_20.4017	Col156	1,1	10.524	0,46	✗	✓	0,6	0	0
NODE_47_length_5294_cov_360.725	-	19,2	5.294	0,48	✗	✓	0,7	0	0
NODE_51_length_4746_cov_19.3334	-	1	4.746	0,58	✗	✓	27,3	0	0
NODE_53_length_4196_cov_304.85	-	16,2	4.196	0,5	✗	✓	0,6	0	0
NODE_54_length_3814_cov_20.5319	-	1,1	3.814	0,52	✗	✓	7,2	0	0

Figure 8: Integration of the Platon analysis workflow results into the ASA³P reports.

Detected plasmid contigs are comprehensively characterized via Platon. Results are integrated into and presented as detailed ASA³P reports.

Its non-interactive workflow and robust classification performance makes Platon a suitable fit and ideal for the integration into ASA³P for the automated detection and characterization of plasmid fragments from bacterial draft assemblies in a taxonomy-independent manner [441]. At the time of writing, Platon has been integrated into ASA³P and is currently undergoing testing and debugging. Results of Platon's analysis workflow are part of ASA³P's interactive HTML reports (Figure 8). To additionally indicate public plasmid sequences that are potentially contained within given draft genomes, visualizations of possible plasmid sequence reconstructions are integrated into ASA³P's reports (Figure 9). At the time of writing, Platon has been cited 19 times due to Dimensions.ai.

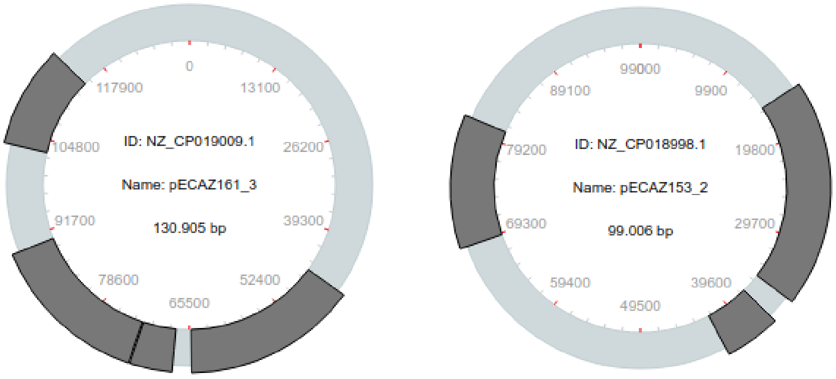


Figure 9: Visualization of potential plasmid reconstructions via detected plasmid contigs.

Plasmid contigs of a bacterial draft assembly were detected with Platon and mapped onto complete reference plasmids. Reference plasmids with contig hits of two or more contigs are visualized along with contig hits in order to reveal potential reconstructions of plasmids within a certain genome. Light gray circle: reference plasmid sequence; dark gray regions: alignments of contigs resulting from short-read draft assemblies.

4.2.2 Ongoing developments and potential improvements of Platon

An interesting aspect for future explorations that has not yet been addressed is the application of Platon for the analysis of whole-metagenome sequencing data. Initially, Platon has been developed to detect plasmid-borne contigs from WGS draft assemblies of single bacterial isolates. However, as the underlying methodology and its implementation is not fundamentally bound to assemblies of isolated genomes, in principle, the methodology is also applicable to metagenomic approaches. However, due to the significantly larger data sizes implied, it might be beneficial or even necessary to analyze assembled metagenomic contigs in parallel. This is already implemented for the various contig characterization steps but not yet for the *ab initio* prediction of protein-coding genes as well as the lookup of MPS. A reimplementation in a dedicated workflow engine providing divide-and-conquer approaches for large numbers of contigs, as for instance NextFlow [442], will significantly improve vertical scalability, add new horizontal scalability features and thus, will reduce the overall runtime of the software.

Furthermore, while debugging and benchmarking Platon it became obvious that many contigs are located within diffuse niches of the multidimensional feature space reflecting the various contig characterizations. These contigs are hard to classify via simple heuristics. Hence, it might be rewarding to address this challenging classification task by using machine-learning approaches, *e.g.* artificial neural networks, to target these partially non-linear properties of the data. These approaches might utilize, combine and thus take advantage of the various general and plasmid specific features, as for instance differences in the GC content, contig lengths, coverages and k-mer frequencies. A recent study has shown that machine-learning approaches utilizing combinations of these features are able to provide competitive results [443].

An idea for future investigation directly aims at Platon's underlying RDS methodology that takes into account differences of protein sequence homology search hits on complete replicons, *i.e.* chromosomes and plasmids. This binary homology search could easily be expanded to other sequence types and thus generalized for the detection of further mobile genetic elements harboring protein-coding genes like prophages. In principle, this approach can be used on all protein encoding DNA sequences that can be separated into two or more categories. It might be rewarding to reuse and test this approach for other detection or classification tasks.

Another example for future enhancements follows up on the visual indication of publicly known plasmid sequences that are potentially contained in assembled draft genomes (Figure 8). Instead of the mere visual indication of public plasmid sequences taking into account contigs that have priorly been identified to be plasmid-borne on their own, databases of known plasmid sequences could be screened for potential reconstructions based on all contigs that are present in a given draft assembly. In addition, this would facilitate the parameterized screening for plasmid sequences allowing for either relaxed or conservative searches.

4.3 Rapid and automated determination of suitable reference genomes with ReferenceSeeker

"Taxonomy is described
sometimes as a science and sometimes as an art,
but really it's a battleground."

Bill Bryson
A Short History of Nearly Everything

Selecting suitable microbial reference genomes is a necessary task for many WGS data analyses. Due to the large and constantly rising number of publicly available genomes, this selection process becomes more and more difficult. Many contemporary bioinformatics software tools are provided via online services [274,275] or interactive graphical user interface (GUI) implementations [274] and thus, are not applicable to large-scale data analysis. However, available tools that are locally executable via a command line interface, do not provide integrated databases [276] or do not achieve sufficient resolutions at the required strain level [374,377]. Likewise, some tools are not usable for more-distantly related genomes [276]. To solve this issue, this thesis provides a new bioinformatic software tool called ReferenceSeeker for the scalable command line search for suitable microbial reference genomes from large integrated databases [444]. To achieve this task, the implemented two-step approach combines a rapid k-mer fingerprint lookup of potential reference genome candidates with the robust and thorough calculation of ANI and conserved DNA values. It scales vertically and thus achieves short wall-clock runtimes. ANI based genome to genome distances allow reasonably detailed comparisons of query and reference genomes, even at small DNA fragment levels. Default values for ANI and conserved DNA thresholds are set to well known boundaries for bacterial species. However, these are adjustable parameters allowing more or less constrained taxonomical searches to increase potential applications [273].

In contrast to existing tools, ReferenceSeeker provides a dedicated database integrating k-mer fingerprints, taxonomic information as well as compressed DNA sequences of all entries in the reference genome database. For further convenience, five pre-compiled databases are provided for the following microbial taxa: viruses, archaea, bacteria, fungi and protozoa. Noteworthy, as the implemented approach is generally able to compare all types of larger DNA sequences, recently, a dedicated plasmid database comprising 26,907 sequences has been compiled and publicly provided via Zenodo [445]. Furthermore, the software provides a command line interface for streamlined compilations of custom databases and the local import of available genomes or DNA sequences. Thus, users are able to create dedicated customized local databases for targeted taxonomic analysis. At the time of writing, ReferenceSeeker has been cited two times due to Dimensions.ai.

Although this tool was developed for the automated lookup of suitable reference genomes, it is also a useful tool for taxonomic classifications of bacterial genomes by applying generally accepted thresholds for ANI and conserved DNA values [276].

4.3.1 Integration into ASA³P

The integrated and fast implementation makes ReferenceSeeker another appropriate fit for the close integration into ASA³P. In the recent release v1.3.0 after the initial publication, the k-mer based taxonomic classification using Kraken [439] was replaced by ReferenceSeeker taking advantage of the ANI methodology for which widely accepted species boundary thresholds exist. This replacement significantly reduced ASA³P's storage requirements from 142 GB to only 29 GB and thus increased its general usability, especially for installations on standard consumer hardware providing only limited hardware capacities.

4.3.2 Ongoing developments

ReferenceSeeker and other contemporary bioinformatics software tools for the calculation of inter-genomic distances are currently only applicable for the automated lookup of reference genomes based on a single query genome. To analyze, for instance, a cohort of bacterial genomes, this single reference genome should ideally reflect as much as possible of the entire genomic landscape of all genomes in this analysis, in order not to unintentionally mask certain genomic regions in a SNP calling analysis. A further application requiring sufficiently related syntenic genomic regions is the ordering and rearrangement of contigs within draft assemblies. However, the larger a group of genomes or the more diverse its members, the more demanding this selection becomes. Currently, ASA³P users are required to provide at least one, better more, closely related reference genomes. Of course, these manual selections of reference genomes introduce an unnecessary bias as it remains questionable if manually selected reference genomes always reflect the optimal choice from the hundreds and thousands of available public genomes.

To address these issues and to fully automate this selection process for subsequent bacterial cohort analyses with ASA³P, the expansion of the ReferenceSeeker workflow from the current 1:*n* to an *m*:*n* approach for query and reference genomes, respectively, is a promising approach, which is currently being addressed in an ongoing Master thesis at the time of writing.

5 Conclusion

The giant progress in DNA sequencing technologies revolutionized the field of microbial genomics. Vast numbers of genomes are sequenced worldwide every day and many research areas benefit tremendously from these developments, in particular medical microbiology and epidemiology. Nowadays, genome-based analyses are essential tools for the detection, classification, typing and comparison of special-interest genes and genomes at various levels. At the same time, IT is revolutionized alike by new trends such as software containerization and cloud computing. New software engineering paradigms and frameworks have recently emerged to conduct robust and scalable computations executed on distributed and heterogeneous IT infrastructures. Albeit the mere sequencing of bacterial genomes as well as computing capacity in general are not limiting factors anymore, the comprehensive, timely and standardized analysis of all this data however remains an issue of rising importance.

This thesis provides novel bioinformatics software tools for the fully automated and scalable analysis of WGS data of small and large cohorts of bacterial genomes. As a first contribution, ASA³P directly addresses this objective. In contrast to existing software tools, it offers a unique and comprehensive combination of features in terms of support for different DNA sequencing platforms and assembly approaches, thorough per-isolate characterization, comparative analyses, and both vertical and horizontal scalability. It supports researchers with a single software suite for the collective analysis of bacterial genomes and furthermore allows the seamless upscaling from small to vast numbers of genomes using regular consumer hardware or HPC and CCIs, respectively. A second and third contribution comprise a novel bioinformatic methodology and two new software tools addressing distinct issues that have arised from the design of this workflow. To improve the integrated analysis of plasmids, RDSs were introduced as a new approach for the automated and taxonomy-independent detection of plasmid-borne contigs from draft assemblies. It achieves a robust and balanced classification performance and was implemented in Platon. To streamline both the automated selection of closely related reference genomes and the taxonomic classification of assembled genomes, a novel approach combining existing tools and methodologies has been implemented in

ReferenceSeeker. Due to their automated taxonomy-independent workflows and integrated databases, these tools fit both the scientific and technical requirements for the integration into ASA³P. Furthermore, they are available as stand-alone bioinformatic software tools, as well.

These contributions have already been used in various studies and publications. In particular, ASA³P has been shown to be a useful tool for researchers in the field of medical microbiology and epidemiology streamlining the data processing and genome characterization workflow. It enables researchers to take advantage of scalable IT resources and a diverse set of proven bioinformatics software tools dedicated to the various tasks involved. Hence, even more bacterial genomes and larger cohorts thereof can be analyzed, characterized and compared, allowing to keep up with DNA sequencing technologies and future demands. It will help to address urgent issues in the field of medical microbiology as for instance, AMR and the spread of pathogenic bacteria. However, it must be mentioned that by no means ASA³P is restricted to these applications. The robust and extensible framework of this software provides a platform that can be expanded and adapted. Hence, many research areas that include the analysis of bacterial genomes, *e.g.* biotechnology, veterinary medicine, microbial ecology and space microbiology, might benefit from these automated and scalable solutions opening further applications within the much larger research field of microbial genomics. Furthermore, new questions and ideas for improvements and potential new tools emerged from this thesis regarding for example, reference genomes for entire cohorts, the improvement of contig classifications via machine learning approaches and potential metagenome applications, the screening of publicly known plasmid sequences and the annotation of bacterial genomes. These ideas provide interesting and promising subjects for further investigations and future research projects.

6 References

1. Campbell NA, Reece JB. Biology. Benjamin Cummings; 2002.

2. Fraser CM, Eisen JA, Salzberg SL. Microbial genome sequencing. Nature. 2000;406: 799–803. doi:10.1038/35021244

3. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH Human Microbiome Project. Genome Res. 2009;19: 2317–2323. doi:10.1101/gr.096651.109

4. Relman DA, Falkow S. The meaning and impact of the human genome sequence for microbiology. Trends Microbiol. 2001;9: 206–208. doi:10.1016/s0966-842x(01)02041-8

5. Relman DA. New technologies, human-microbe interactions, and the search for previously unrecognized pathogens. J Infect Dis. 2002;186 Suppl 2: S254–8. doi:10.1086/344935

6. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14: e1002533. doi:10.1371/journal.pbio.1002533

7. Appanna VD. Human Microbes - The Power Within: Health, Healing and Beyond. Springer; 2018.

8. Khan R, Petersen FC, Shekhar S. Commensal Bacteria: An Emerging Player in Defense Against Respiratory Pathogens. Front Immunol. 2019;10: 1203. doi:10.3389/fimmu.2019.01203

9. Abt MC, Pamer EG. Commensal bacteria mediated defenses against pathogens. Curr Opin Immunol. 2014;29: 16–22. doi:10.1016/j.coi.2014.03.003

10. Man WH, de Steenhuijsen Piters WAA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. Nat Rev Microbiol. 2017;15: 259–270. doi:10.1038/nrmicro.2017.14

11. Kim WJ, Higashi D, Goytia M, Rendón MA, Pilligua-Lucas M, Bronnimann M, et al. Commensal *Neisseria* Kill *Neisseria gonorrhoeae* through a DNA-Dependent Mechanism. Cell Host Microbe. 2019;26: 228–239.e8. doi:10.1016/j.chom.2019.07.003

12. Green MH. Editor's Introduction to Pandemic Disease in the Medieval World: Rethinking the Black Death. The Medieval Globe. 2014;1: 3

13. Harbeck M, Seifert L, Hänsch S, Wagner DM, Birdsell D, Parise KL, et al. Yersinia pestis DNA from Skeletal Remains from the 6th Century AD Reveals Insights into Justinianic Plague. PLoS Pathogens. 2013. p. e1003349. doi:10.1371/journal.ppat.1003349

14. Alcon SA. A Pest In The Land: New World Epidemics In A Global Perspective, Albuquerque. University of New Mexico Press; 2003.

15. Wheelis M. Biological warfare at the 1346 siege of Caffa. Emerg Infect Dis. 2002;8: 971–975. doi:10.3201/eid0809.010536

16. Randremanana R, Andrianaivoarimanana V, Nikolay B, Ramasindrazana B, Paireau J, ten Bosch QA, et al. Epidemiological characteristics of an urban plague epidemic in Madagascar, August–November, 2017: an outbreak report. Lancet Infect Dis. 2019;19: 537–545. doi:10.1016/S1473-3099(18)30730-8

17. Cabanel N, Leclercq A, Chenal-Francisque V, Annajar B, Rajerison M, Bekkhoucha S, et al. Plague outbreak in Libya, 2009, unrelated to plague in Algeria. Emerg Infect Dis. 2013;19: 230–236. doi:10.3201/eid1902.121031

18. Papagrigorakis MJ, Synodinos PN, Yapijakis C. Ancient typhoid epidemic reveals possible ancestral strain of *Salmonella enterica* serovar Typhi. Infect Genet Evol. 2007;7: 126–127. doi:10.1016/j.meegid.2006.04.006

19. Crump JA, Luby SP, Mintz ED. The global burden of typhoid fever. Bull World Health Organ. 2004;82: 346–353. Available: https://www.ncbi.nlm.nih.gov/pubmed/15298225

20. Diseases, Vaccine-Preventable: Typhoid and other invasive salmonellosis. World Health Organization; 5, September, 2018. Available: https://www.who.int/immunization/monitoring_surveillance/burden/vpd/WHO_SurveillanceVaccinePreventable_21_Typhoid_R2.pdf

21. Koch R. Die Ätiologie der Tuberkulose. Berliner Klinischen Wochenschrift. 1882;19.

22. Global tuberculosis report 2019. World Health Organization; 2019 Oct. Available: https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1

23. Weekly epidemiological record. World Health Organization; 2018 Sep. Report No.: 38. Available: https://apps.who.int/iris/bitstream/handle/10665/274654/WER9338.pdf?ua=1

24. WHO ESTIMATES OF THE GLOBAL BURDEN OF FOODBORNE DISEASES. World Health Organization; 2015 Dec.

25. Zhi C, Lv L, Yu L-F, Doi Y, Liu J-H. Dissemination of the mcr-1 colistin resistance gene. The Lancet Infectious Diseases. 2016. pp. 292–293. doi:10.1016/s1473-3099(16)00063-3

26. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis. 2016;16: 161–168. doi:10.1016/S1473-3099(15)00424-7

27. Falgenhauer L. Highly conserved plasmids drive the spread of the mobile colistin resistance gene mcr-1 in Germany and Spain. 2017. doi:10.26226/morressier.5991c407d462b80292388e07

28. Falgenhauer L, Schwengers O, Schmiedel J, Baars C, Lambrecht O, Heß S, et al. Multidrug-Resistant and Clinically Relevant Gram-Negative Bacteria Are Present in German Surface Waters. Front Microbiol. 2019;10: 2779. doi:10.3389/fmicb.2019.02779

29. Culliton BJ. Emerging viruses, emerging threat. Science. 1990;247: 279–280. doi:10.1126/science.2153314

30. O'Neill J. The Review on Antimicrobial Resistance Tackling Drug-Resistant Infections Globally: final report and recommendations. Wellcome Trust, UK Department of Health; 2016 May. Available: https://amr-review.org

31. Maragakis LL, Perencevich EN, Cosgrove SE. Clinical and economic burden of antimicrobial resistance. Expert Rev Anti Infect Ther. 2008;6: 751–763. doi:10.1586/14787210.6.5.751

32. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet Infect Dis. 2019;19: 56–66. doi:10.1016/S1473-3099(18)30605-4

33. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, et al. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J Antimicrob Chemother. 2013;68: 771–777. doi:10.1093/jac/dks496

34. van Belkum A, Burnham C-AD, Rossen JWA, Mallard F, Rochas O, Dunne WM Jr. Innovative and rapid antimicrobial susceptibility testing systems. Nat Rev Microbiol. 2020;18: 299–311. doi:10.1038/s41579-020-0327-x

35. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol. 2017;243: 16–24. doi:10.1016/j.jbiotec.2016.12.022

36. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. J Antimicrob Chemother. 2017;72: 1937–1947. doi:10.1093/jac/dkx067

37. Surveillance of antimicrobial resistance in Europe 2018. ECDC; 2019 Nov.

38. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. doi:10.1038/sdata.2016.18

39. Kwong JC, Schultz MB, Williamson DA, Stinear TP, Seemann T, Howden BP. Comment on: Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. The Journal of antimicrobial chemotherapy. 2017. pp. 635–636. doi:10.1093/jac/dkw473

40. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13: e1002195. doi:10.1371/journal.pbio.1002195

41. Empowering App Development for Developers | Docker. In: Docker [Internet]. [cited 9 Jul 2020]. Available: https://www.docker.com/

42. podman.io. [cited 28 Aug 2020]. Available: https://podman.io/

43. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15: 475–476. doi:10.1038/s41592-018-0046-7

44. Porta M. A Dictionary of Public Health. 2nd ed. Last JM, editor. Oxford University Press; 2018. doi:10.1093/acref/9780191844386.001.0001

45. Hippocrates., Lloyd GER, Chadwick J, Mann WN. Hippocratic writings. Harmondsworth; New York: Penguin; 1983. Available: https://www.worldcat.org/title/hippocratic-writings/oclc/10501704

46. Curtis VA. Dirt, disgust and disease: a natural history of hygiene. J Epidemiol Community Health. 2007;61: 660–664. doi:10.1136/jech.2007.062380

47. Halliday S. Death and miasma in Victorian London: an obstinate belief. BMJ. 2001;323: 1469–1471. doi:10.1136/bmj.323.7327.1469

48. Paneth N, Vinten-Johansen P, Brody H, Rip M. A rivalry of foulness: official and unofficial investigations of the London cholera epidemic of 1854. Am J Public Health. 1998;88: 1545–1553. doi:10.2105/ajph.88.10.1545

49. Snow J. On the Mode of Communication of Cholera. John Churchill; 1855.

50. Tapia Granados JA. William Farr. Encyclopædia Britannica. Encyclopædia Britannica, inc.; 2020. Available: https://www.britannica.com/biography/William-Farr

51. Byrne JP. Encyclopedia of the Black Death. ABC-CLIO; 2012.

52. Nutton V. The reception of Fracastoro's Theory of contagion: the seed that fell among thorns? Osiris. 1990;6: 196–134. doi:10.1086/368701

53. Koch R. Die Ätiologie der Milzbrand-Krankheit, begründet auf die Entwicklungsgeschichte des Bacillus Anthracis (1876). In: Gradmann C, editor. Robert Koch : Zentrale Texte. Berlin, Heidelberg: Springer Berlin Heidelberg; 2018. pp. 19–43. doi:10.1007/978-3-662-56454-7_2

54. Deutsche medizinische Wochenschrift. Georg Thieme Verlag.; 1884.

55. Vouga M, Greub G. Emerging bacterial pathogens: the past and beyond. Clin Microbiol Infect. 2016;22: 12–21. doi:10.1016/j.cmi.2015.10.010

56. Approved lists of bacterial names. Med J Aust. 1980;2: 3–4. doi:10.1099/00207713-30-1-225

57. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, et al. Microbiology in the post-genomic era. Nat Rev Microbiol. 2008;6: 419–430. doi:10.1038/nrmicro1901

58. Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell. 1984;37: 67–75. doi:10.1016/0092-8674(84)90301-5

59. PulseNet Timeline | PulseNet | CDC. 2 May 2019 [cited 2 Jun 2020]. Available: https://www.cdc.gov/pulsenet/anniversary/timeline.html

60. Fast Facts about PulseNet | About | PulseNet | CDC. 27 Feb 2019 [cited 2 Jun 2020]. Available: https://www.cdc.gov/pulsenet/about/fast-facts.html

61. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95: 3140–3145. doi:10.1073/pnas.95.6.3140

62. Nicolas P, Raphenon G, Guibourdenche M, Decousset L, Stor R, Gaye AB. The 1998 Senegal epidemic of meningitis was due to the clonal expansion of A:4:P1.9, clone III-1, sequence type 5 *Neisseria meningitidis* strains. J Clin Microbiol. 2000;38: 198–200. doi:10.1128/JCM.38.1.198-200.2000

63. Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J Clin Microbiol. 2000;38: 1008–1015. doi:10.1128/JCM.38.3.1008-1015.2000

64. Feil EJ, Smith JM, Enright MC, Spratt BG. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics. 2000;154: 1439–1450. doi:10.1093/genetics/154.4.1439

65. Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, et al. Multilocus sequence typing system for *Campylobacter jejuni*. J Clin Microbiol. 2001;39: 14–23. doi:10.1128/JCM.39.1.14-23.2001

66. Kotetishvili M, Stine OC, Kreger A, Morris JG Jr, Sulakvelidze A. Multilocus sequence typing for characterization of clinical and environmental salmonella strains. J Clin Microbiol. 2002;40: 1626–1635. doi:10.1128/JCM.40.5.1626-1635.2002

67. Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E, et al. Multilocus sequence typing scheme for Enterococcus faecium. J Clin Microbiol. 2002;40: 1963–1971. doi:10.1128/JCM.40.6.1963-1971.2002

68. van Loo IHM, Heuvelman KJ, King AJ, Mooi FR. Multilocus sequence typing of *Bordetella pertussis* based on surface protein genes. J Clin Microbiol. 2002;40: 1994–2001. doi:10.1128/JCM.40.6.1994-2001.2002

69. Chan MS, Maiden MC, Spratt BG. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics. 2001;17: 1077–1083. doi:10.1093/bioinformatics/17.11.1077

70. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11: 595. doi:10.1186/1471-2105-11-595

71. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

72. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med. 2011;364: 33–42. doi:10.1056/NEJMoa1012928

73. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, et al. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clin Infect Dis. 2016;63: 380–386. doi:10.1093/cid/ciw242

74. Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. Genome Res. 2013;23: 653–664. doi:10.1101/gr.147710.112

75. Pérez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. Infect Genet Evol. 2018;63: 346–359. doi:10.1016/j.meegid.2017.09.022

76. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, et al. Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*. J Clin Microbiol. 2016;54: 333–342. doi:10.1128/JCM.02344-15

77. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen Genomics in Public Health. N Engl J Med. 2019;381: 2569–2580. doi:10.1056/NEJMsr1813907

78. Scharff RL, Besser J, Sharp DJ, Jones TF, Peter G-S, Hedberg CW. An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. Am J Prev Med. 2016;50: S66–S73. doi:10.1016/j.amepre.2015.09.018

79. Kubota KA, Wolfgang WJ, Baker DJ, Boxrud D, Turner L, Trees E, et al. PulseNet and the Changing Paradigm of Laboratory-Based Surveillance for Foodborne Diseases. Public Health Rep. 2019;134: 22S–28S. doi:10.1177/0033354919881650

80. Ribot EM, Hise KB. Future challenges for tracking foodborne diseases: PulseNet, a 20-year-old US surveillance system for foodborne diseases, is expanding both globally and technologically. EMBO Rep. 2016;17: 1499–1505. doi:10.15252/embr.201643128

81. Taboada EN, Graham MR, Carriço JA, Van Domselaar G. Food Safety in the Age of Next Generation Sequencing, Bioinformatics, and Open Data Access. Front Microbiol. 2017;8: 909. doi:10.3389/fmicb.2017.00909

82. Oniciuc EA, Likotrafiti E, Alvarez-Molina A, Prieto M, Santos JA, Alvarez-Ordóñez A. The Present and Future of Whole Genome Sequencing (WGS) and Whole Metagenome Sequencing (WMS) for Surveillance of Antimicrobial Resistant Microorganisms and Antimicrobial Resistance Genes across the Food Chain. Genes . 2018;9. doi:10.3390/genes9050268

83. Allard M, Wang C, Kastanis G, Pirone C, Muruvanda T, Strain E, et al. Genometrakr: A pathogen database to build a global genomic network for pathogen traceback and outbreak detection. 2015 Annual Meeting. 2015.

84. Timme RE, Sanchez Leon M, Allard MW. Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. Methods Mol Biol. 2019;1918: 201–212. doi:10.1007/978-1-4939-9000-9_17

85. Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Agama Study Group, Achtman M. The EnteroBase user's guide, with case studies on *Salmonella transmissions*, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. Genome Res. 2020;30: 138–152. doi:10.1101/gr.251678.119

86. The Centre for Genomic Pathogen Surveillance. Pathogenwatch | A Global Platform for Genomic Surveillance. [cited 26 Aug 2020]. Available: https://pathogen.watch

87. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. Nat Rev Microbiol. 2020;18: 344–359. doi:10.1038/s41579-019-0315-1

88. Acman M, van Dorp L, Santini JM, Balloux F. Large-scale network analysis captures biological features of bacterial plasmids. Nat Commun. 2020;11: 1–11. doi:10.1038/s41467-020-16282-w

89. Bobay L-M. The Prokaryotic Species Concept and Challenges. In: Tettelin H, Medini D, editors. The Pangenome: Diversity, Dynamics and Evolution of Genomes. Cham: Springer International Publishing; 2020. pp. 21–49. doi:10.1007/978-3-030-38281-0_2

90. Lees JA, Tien Mai T, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, et al. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. MBio. 2020;11. doi:10.1128/mBio.01344-20

91. Rocha EPC. The organization of the bacterial genome. Annu Rev Genet. 2008;42: 211–233. doi:10.1146/annurev.genet.42.110807.091653

92. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15: 141–161. doi:10.1007/s10142-015-0433-4

93. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat Rev Microbiol. 2021. doi:10.1038/s41579-020-00497-1

94. Pilla G, Tang CM. Going around in circles: virulence plasmids in enteric pathogens. Nat Rev Microbiol. 2018;16: 484–495. doi:10.1038/s41579-018-0031-2

95. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020. doi:10.1038/s41587-020-0501-8

96. Falush D. Toward the use of genomics to study microevolutionary change in bacteria. PLoS Genet. 2009;5: e1000627. doi:10.1371/journal.pgen.1000627

97. Field D, Hughes J, Moxon ER. Using the genome to understand pathogenicity. Methods Mol Biol. 2004;266: 261–287. doi:10.1385/1-59259-763-7:261

98. Burrack LS, Higgins DE. Genomic approaches to understanding bacterial virulence. Curr Opin Microbiol. 2007;10: 4–9. doi:10.1016/j.mib.2006.11.004

99. Wren BW. Microbial genome analysis: insights into virulence, host adaptation and evolution. Nat Rev Genet. 2000;1: 30–39. doi:10.1038/35049551

100. Baker S, Thomson N, Weill F-X, Holt KE. Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. Science. 2018;360: 733–738. doi:10.1126/science.aar3777

101. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. Elife. 2020;9. doi:10.7554/eLife.53886

102. Griffith F. The Significance of Pneumococcal Types. J Hyg . 1928;27: 113–159. doi:10.1017/s0022172400031879

103. Avery OT, Macleod CM, McCarty M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. J Exp Med. 1944;79: 137–158. doi:10.1084/jem.79.2.137

104. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol. 1952;36: 39–56. doi:10.1085/jgp.36.1.39

105. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171: 737–738. doi:10.1038/171737a0

106. Beadle GW, Tatum EL. Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci U S A. 1941;27: 499–506. doi:10.1073/pnas.27.11.499

107. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3: 318–356. doi:10.1016/s0022-2836(61)80072-7

108. Khorana HG. Polynucleotide synthesis and the genetic code. Harvey Lect. 1966;62: 79–105. Available: https://www.ncbi.nlm.nih.gov/pubmed/4875306

109. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94: 441–448. doi:10.1016/0022-2836(75)90213-2

110. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74: 560–564. doi:10.1073/pnas.74.2.560

111. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265: 687–695. doi:10.1038/265687a0

112. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467. Available: https://www.ncbi.nlm.nih.gov/pubmed/271968

113. The Nobel Prize in Chemistry 1980. In: NobelPrize.org [Internet]. [cited 8 Jun 2020]. Available: https://www.nobelprize.org/prizes/chemistry/1980/sanger/biographical/

114. Staden R. Sequence data handling by computer. Nucleic Acids Res. 1977;4: 4037–4051. doi:10.1093/nar/4.11.4037

115. Staden R. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 1979;6: 2601–2610. doi:10.1093/nar/6.7.2601

116. Sanger F, Coulson AR, Barrell BG, Smith AJ, Roe BA. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J Mol Biol. 1980;143: 161–178. doi:10.1016/0022-2836(80)90196-5

117. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. 1981;9: 3015–3027. doi:10.1093/nar/9.13.3015

118. Schneider TD, Stormo GD, Haemer JS, Gold L. A design for computer nucleic-acid-sequence storage, retrieval, and manipulation. Nucleic Acids Res. 1982;10: 3013–3024. doi:10.1093/nar/10.9.3013

119. Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, et al. The GenBank genetic sequence databank. Nucleic Acids Res. 1986;14: 1–4. doi:10.1093/nar/14.1.1

120. Hamm GH, Cameron GN. The EMBL data library. Nucleic Acids Res. 1986;14: 5–9. doi:10.1093/nar/14.1.5

121. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science. 1987;238: 336–341. doi:10.1126/science.2443975

122. Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res. 1985;13: 2399–2412. doi:10.1093/nar/13.7.2399

123. Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res. 1987;15: 4593–4602. doi:10.1093/nar/15.11.4593

124. Swerdlow H, Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 1990;18: 1415–1419. doi:10.1093/nar/18.6.1415

125. Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, et al. High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res. 1990;18: 4417–4421. doi:10.1093/nar/18.15.4417

126. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227: 1435–1441. doi:10.1126/science.2983426

127. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85: 2444–2448. doi:10.1073/pnas.85.8.2444

128. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

129. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. J Mol Biol. 1982;162: 729–773. doi:10.1016/0022-2836(82)90546-0

130. Miller MJ, Powell JI. A quantitative comparison of DNA sequence assembly programs. J Comput Biol. 1994;1: 257–269. doi:10.1089/cmb.1994.1.257

131. Gryan G. Faster sequence assembly software for megabase shotgun assemblies. Genome Sequencing and Analysis Conference VI. 1994.

132. Sutton GG, White O, Adams MD, Kerlavage AR. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Science and Technology. 1995;1: 9–19. doi:10.1089/gst.1995.1.9

133. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269: 496–512. doi:10.1126/science.7542800

134. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of *Mycoplasma genitalium*. Science. 1995;270: 397–403. doi:10.1126/science.270.5235.397

135. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 1996;24: 4420–4449. doi:10.1093/nar/24.22.4420

136. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of *Escherichia coli* K-12. Science. 1997;277: 1453–1462. doi:10.1126/science.277.5331.1453

137. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature. 1997;390: 249–256. doi:10.1038/36786

138. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature. 1997;388: 539–547. doi:10.1038/41483

139. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature. 1997;390: 580–586. doi:10.1038/37551

140. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science. 1998;281: 375–388. doi:10.1126/science.281.5375.375

141. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393: 537–544. doi:10.1038/31159

142. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature. 1999;397: 176–180. doi:10.1038/16495

143. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, et al. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet. 2001;357: 1225–1240. doi:10.1016/s0140-6736(00)04403-2

144. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999;27: 4636–4641. doi:10.1093/nar/27.23.4636

145. Besemer J, Borodovsky M. Heuristic approach to deriving models for gene finding. Nucleic Acids Res. 1999;27: 3911–3920. doi:10.1093/nar/27.19.3911

146. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29: 2607–2618. doi:10.1093/nar/29.12.2607

147. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25: 955–964. doi:10.1093/nar/25.5.955

148. Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M. EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res. 1998;26: 50–53. doi:10.1093/nar/26.1.50

149. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30. doi:10.1093/nar/28.1.27

150. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, et al. GenDB--an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 2003;31: 2187–2195. Available: https://www.ncbi.nlm.nih.gov/pubmed/12682369

151. Bryson K, Loux V, Bossy R, Nicolas P, Chaillou S, van de Guchte M, et al. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. 2006;34: 3533–3545. doi:10.1093/nar/gkl471

152. Bailey LC Jr, Fischer S, Schug J, Crabtree J, Gibson M, Overton GC. GAIA: framework annotation of genomic sequence. Genome Res. 1998;8: 234–250. doi:10.1101/gr.8.3.234

153. Hyman ED. A new method of sequencing DNA. Anal Biochem. 1988;174: 423–436. doi:10.1016/0003-2697(88)90041-3

154. Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity. Anal Biochem. 1987;167: 235–238. doi:10.1016/0003-2697(87)90158-8

155. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281: 363, 365. doi:10.1126/science.281.5375.363

156. McKernan K, Blanchard A, Kotler L, Costa G. Reagents, methods, and libraries for bead-based sequencing. US Patent. 20080003571:A1, 2008. Available: https://patentimages.storage.googleapis.com/23/3d/db/17c00cb479de72/US20080003571A1.pdf

157. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A. 2003;100: 8817–8822. doi:10.1073/pnas.1133470100

158. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006;34: e22. doi:10.1093/nar/gnj023

159. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55: 641–658. doi:10.1373/clinchem.2008.112789

160. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008. pp. 872–876. doi:10.1038/nature06884

161. FTC Challenges Illumina's Proposed Acquisition of PacBio. 17 Dec 2019 [cited 8 Oct 2020]. Available: https://www.ftc.gov/news-events/press-releases/2019/12/ftc-challenges-illuminas-proposed-acquisition-pacbio

162. Beioley K. Illumina and Pacific Biosciences call off merger. Financial Times. 3 Jan 2020. Available: https://www.ft.com/content/01758520-2e38-11ea-bc77-65e4aa615551. Accessed 8 Oct 2020.

163. Illumina Acquires Enancio's Compression Software. [cited 16 Jul 2020]. Available: https://www.illumina.com/company/news-center/feature-articles/illumina-acquires-enancio-s-compression-software.html

164. The NovaSeq 6000 Sequencing System Specification Sheet | Illumina. [cited 8 Jun 2020]. Available: https://science-docs.illumina.com/documents/Instruments/novaseq-6000-spec-sheet-html-770-2016-025/Content/Source/Instruments/NovaSeq/novaseq-6000-spec-sheet-770-2016-025/novaseq-system-spec-sheet-html-770-2016-025.html

165. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). In: National Human Genome Research Institute [Internet]. [cited 16 Jul 2020]. Available: https://www.genome.gov/sequencingcostsdata

166. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

167. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890. doi:10.1093/bioinformatics/bty560

168. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

169. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27: 863–864. doi:10.1093/bioinformatics/btr026

170. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

171. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018;19: 153. doi:10.1186/s13059-018-1540-z

172. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20: 265–272. doi:10.1101/gr.097261.109

173. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–829. doi:10.1101/gr.074492.107

174. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19: 1117–1123. doi:10.1101/gr.089532.108

175. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008;18: 810–820. doi:10.1101/gr.7337908

176. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9: e112963. doi:10.1371/journal.pone.0112963

177. Oxford Nanopore Technologies. Medaka. 2020. Available: https://github.com/nanoporetech/medaka

178. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14: 1394–1403. doi:10.1101/gr.2289704

179. Kolmogorov M, Raney B, Paten B, Pham S. Ragout-a reference-assisted assembly tool for bacterial genomes. Bioinformatics. 2014;30: i302–9. doi:10.1093/bioinformatics/btu280

180. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, et al. MeDuSa: A multi-draft based scaffolder. Bioinformatics. 2015;31: 2443–2451. doi:10.1093/bioinformatics/btv171

181. Chen K-T, Chen C-J, Shen H-T, Liu C-L, Huang S-H, Lu CL. Multi-CAR: a tool of contig scaffolding using multiple references. BMC Bioinformatics. 2016;17: 469. doi:10.1186/s12859-016-1328-7

182. Chen K-T, Shen H-T, Lu CL. Multi-CSAR: a multiple reference-based contig scaffolder using algebraic rearrangements. BMC Syst Biol. 2018;12: 139. doi:10.1186/s12918-018-0654-y

183. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome assembly of large and complex genomes using multiple references. Genome Res. 2018;28: 1720–1732. doi:10.1101/gr.236273.118

184. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18: 1851–1858. doi:10.1101/gr.078212.108

185. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

186. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25. doi:10.1186/gb-2009-10-3-r25

187. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

188. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25: 1966–1967. doi:10.1093/bioinformatics/btp336

189. Blom J, Jakobi T, Doppmeier D, Jaenicke S, Kalinowski J, Stoye J, et al. Exact and complete short-read alignment to microbial genomes using Graphics Processing Unit programming. Bioinformatics. 2011;27: 1351–1358. doi:10.1093/bioinformatics/btr151

190. Liu C-M, Wong T, Wu E, Luo R, Yiu S-M, Li Y, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics. 2012;28: 878–879. doi:10.1093/bioinformatics/bts061

191. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012;22: 557–567. doi:10.1101/gr.131383.111

192. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, et al. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014;30: 2709–2716. doi:10.1093/bioinformatics/btu391

193. Utturkar SM, Klingeman DM, Hurt RA Jr, Brown SD. A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies. Front Microbiol. 2017;8: 1272. doi:10.3389/fmicb.2017.01272

194. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323: 133–138. doi:10.1126/science.1162986

195. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37: 1155–1162. doi:10.1038/s41587-019-0217-9

196. Howorka S, Cheley S, Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. Nat Biotechnol. 2001;19: 636–639. doi:10.1038/90236

197. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4: 265–270. doi:10.1038/nnano.2009.12

198. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol. 2016;34: 518–524. doi:10.1038/nbt.3423

199. World first: continuous DNA sequence of more than a million bases achieved with nanopore sequencing. In: Oxford Nanopore Technologies [Internet]. 27 Dec 2017 [cited 8 Jun 2020]. Available: https://nanoporetech.com/about-us/news/world-first-continuous-dna-sequence-more-million-bases-achieved-nanopore-sequencing

200. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. 2019;35: 2193–2198. doi:10.1093/bioinformatics/bty841

201. MinION | Oxford Nanopore Technologies. [cited 9 Oct 2020]. Available: https://nanoporetech.com/products/minion

202. The highest throughput yet: PromethION breaks the 7 Terabase mark. In: Oxford Nanopore Technologies [Internet]. 16 Apr 2019 [cited 8 Jun 2020]. Available: https://nanoporetech.com/about-us/news/highest-throughput-yet-promethion-breaks-7-terabase-mark

203. PromethION | Oxford Nanopore Technologies. [cited 9 Oct 2020]. Available: https://nanoporetech.com/products/promethion

204. Van der Verren SE, Van Gerven N, Jonckheere W, Hambley R, Singh P, Kilgour J, et al. A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. Nat Biotechnol. 2020. doi:10.1038/s41587-020-0570-8

205. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019;20: 129. doi:10.1186/s13059-019-1727-y

206. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19: R227–40. doi:10.1093/hmg/ddq416

207. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107: 1–8. doi:10.1016/j.ygeno.2015.11.003

208. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10: 563–569. doi:10.1038/nmeth.2474

209. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13: 1050–1054. doi:10.1038/nmeth.4035

210. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27: 722–736. doi:10.1101/gr.215087.116

211. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27: 737–746. doi:10.1101/gr.214270.116

212. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37: 540–546. doi:10.1038/s41587-019-0072-8

213. Vaser R, Šikić M. Raven: a de novo genome assembler for long reads. 2020. p. 2020.08.07.242461. doi:10.1101/2020.08.07.242461

214. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17: 155–158. doi:10.1038/s41592-019-0669-3

215. Simpson J. Nanopolish. 2020. Available: https://github.com/jts/nanopolish

216. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLoS Comput Biol. 2017;13: e1005595. doi:10.1371/journal.pcbi.1005595

217. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29: 2669–2677. doi:10.1093/bioinformatics/btt476

218. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016;32: 1009–1015. doi:10.1093/bioinformatics/btv688

219. Wick R. Filtlong: quality filtering tool for long reads. Github; Available: https://github.com/rrwick/Filtlong

220. Wick R. Porechop: adapter trimmer for Oxford Nanopore reads. Github; Available: https://github.com/rrwick/Porechop

221. The UniVec Database. [cited 18 Nov 2021]. Available: https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/

222. Longo MS, O'Neill MJ, O'Neill RJ. Abundant human DNA contamination identified in non-primate genome databases. PLoS One. 2011;6: e16410. doi:10.1371/journal.pone.0016410

223. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res. 2018;7: 1338. doi:10.12688/f1000research.15931.2

224. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34: 2666–2669. doi:10.1093/bioinformatics/bty149

225. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science. 2000;287: 2196–2204. doi:10.1126/science.287.5461.2196

226. Pevzner PA, Borodovsky MYu, Mironov AA. Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. J Biomol Struct Dyn. 1989;6: 1013–1026. doi:10.1080/07391102.1989.10506528

227. Idury RM, Waterman MS. A new algorithm for DNA sequence assembly. J Comput Biol. 1995;2: 291–306. doi:10.1089/cmb.1995.2.291

228. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98: 9748–9753. doi:10.1073/pnas.171285098

229. Li H. miniasm: Ultrafast de novo assembly for long noisy reads (though having no consensus step). Github; Available: https://github.com/lh3/miniasm

230. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38: 1044–1053. doi:10.1038/s41587-020-0503-6

231. Vaser R, Šikić M. Yet another de novo genome assembler. 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA). 2019. pp. 147–151. doi:10.1109/ISPA.2019.8868909

232. Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. BMC Genomics. 2020;21: 631. doi:10.1186/s12864-020-07041-8

233. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36: 2253–2255. doi:10.1093/bioinformatics/btz891

234. Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. bioRxiv. 2021. p. 2021.10.14.464465. doi:10.1101/2021.10.14.464465

235. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12: 733–735. doi:10.1038/nmeth.3444

236. Huang N, Nie F, Ni P, Luo F, Gao X, Wang J. NeuralPolish: a novel Nanopore polishing method based on alignment matrix construction and orthogonal Bi-GRU Networks. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab354

237. Oxford Nanopore Technologies Ltd. MeDaKa. Github; Available: https://github.com/nanoporetech/medaka

238. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

239. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from. Cold Spring Harbor Laboratory Press Method. 2015; 1–31. doi:10.1101/gr.186072.114

240. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962: 227–245. doi:10.1007/978-1-4939-9173-0_14

241. Lu CL, Chen K-T, Huang S-Y, Chiu H-T. CAR: contig assembly of prokaryotic draft genomes using rearrangements. BMC Bioinformatics. 2014;15: 381. doi:10.1186/s12859-014-0381-3

242. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

243. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2020. doi:10.1093/nar/gkaa1047

244. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32: 11–16. doi:10.1093/nar/gkh152

245. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49: 9077–9096. doi:10.1093/nar/gkab688

246. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35: 3100–3108. doi:10.1093/nar/gkm160

247. Seemann T. barrnap: BAsic Rapid Ribosomal RNA Predictor. Github; Available: https://github.com/tseemann/barrnap

248. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8: 18. doi:10.1186/1471-2105-8-18

249. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8: 209. doi:10.1186/1471-2105-8-209

250. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One. 2014;9: e110726. doi:10.1371/journal.pone.0110726

251. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics. 2016;17: 356. doi:10.1186/s12864-016-2627-0

252. Wang K, Liang C. CRF: detection of CRISPR arrays using random forest. PeerJ. 2017;5: e3219. doi:10.7717/peerj.3219

253. Padilha VA, Alkhnbashi OS, Shah SA, de Carvalho ACPLF, Backofen R. CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. Gigascience. 2020;9. doi:10.1093/gigascience/giaa062

254. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018;4. doi:10.1099/mgen.0.000206

255. Luo H, Gao F. DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. Nucleic Acids Res. 2019;47: D74–D77. doi:10.1093/nar/gky1014

256. Hyatt D, Chen GL, LoCascio PF. Prodigal: prokaryotic gene recognition and translation initiation site identification. Biomed Chromatogr. 2010. doi:10.1186/1471-2105-11-119

257. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res. 2018;28: 1079–1089. doi:10.1101/gr.230615.117

258. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195

259. Suzuki S, Kakuta M, Ishida T, Akiyama Y. Faster sequence homology searches by clustering subsequences. Bioinformatics. 2015;31: 1183–1190. doi:10.1093/bioinformatics/btu780

260. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028. doi:10.1038/nbt.3988

261. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18: 366–368. doi:10.1038/s41592-021-01101-x

262. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, et al. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 2005;33: W455–9. doi:10.1093/nar/gki593

263. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9: 75. doi:10.1186/1471-2164-9-75

264. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 2018;46: D851–D860. doi:10.1093/nar/gkx1068

265. Dong Y, Li C, Kim K, Cui L, Liu X. Genome annotation of disease-causing microorganisms. Brief Bioinform. 2021;22: 845–854. doi:10.1093/bib/bbab004

266. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153

267. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44: 6614–6624. doi:10.1093/nar/gkw569

268. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34: 1037–1039. doi:10.1093/bioinformatics/btx713

269. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz848

270. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41: D590–6. doi:10.1093/nar/gks1219

271. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12: 635–645. doi:10.1038/nrmicro3330

272. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42: D633–42. doi:10.1093/nar/gkt1244

273. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57: 81–91. doi:10.1099/ijs.0.64483-0

274. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106: 19126–19131. doi:10.1073/pnas.0906412106

275. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14: 60. doi:10.1186/1471-2105-14-60

276. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9: 5114. doi:10.1038/s41467-018-07641-9

277. Jolley KA, Bray JE, Maiden MCJ. A RESTful application programming interface for the PubMLST molecular typing and genome databases. Database . 2017;2017. doi:10.1093/database/bax060

278. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67: 2640–2644. doi:10.1093/jac/dks261

279. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58: 212–220. doi:10.1128/AAC.01310-13

280. Seemann T. abricate: Mass screening of contigs for antimicrobial and virulence genes. Github; Available: https://github.com/tseemann/abricate

281. de Man TJB, Limbago BM. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. mSphere. 2016;1. doi:10.1128/mSphere.00050-15

282. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48: D517–D525. doi:10.1093/nar/gkz935

283. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11: 12728. doi:10.1038/s41598-021-91456-0

284. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Res. 2009;37: D443–7. doi:10.1093/nar/gkn656

285. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015;9: 207–216. doi:10.1038/ismej.2014.106

286. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res. 2017;45: D574–D580. doi:10.1093/nar/gkw1009

287. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017;45: D535–D542. doi:10.1093/nar/gkw1017

288. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 2019;47: D687–D692. doi:10.1093/nar/gky1080

289. Blin K, Pascal Andreu V, de Los Santos ELC, Del Carratore F, Lee SY, Medema MH, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2019;47: D625–D630. doi:10.1093/nar/gky1060

290. Carroll LM, Larralde M, Fleck JS, Ponnudurai R, Milanese A, Cappio E, et al. Accurate de novo identification of biosynthetic gene clusters with GECCO. bioRxiv. 2021. p. 2021.05.03.442509. doi:10.1101/2021.05.03.442509

291. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34: D32–6. doi:10.1093/nar/gkj014

292. Biswas A, Gauthier DT, Ranjan D, Zubair M. ISQuest: Finding insertion sequences in prokaryotic sequence fragment data. Bioinformatics. 2015;31: 3406–3412. doi:10.1093/bioinformatics/btv388

293. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 2017;33: 3340–3347. doi:10.1093/bioinformatics/btx433

294. MGI sequencing platforms: High-throughput gene sequencers, DNBSEQ^{TM} sequencing technology-MGI. [cited 9 Oct 2020]. Available: https://en.mgitech.cn/products/

295. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11: 207. doi:10.1186/gb-2010-11-5-207

296. Fleming A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. Br J Exp Pathol. 1929;10: 226. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048009/

297. Schatz A, Bugle E, Waksman SA. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria.∗†. Proc Soc Exp Biol Med. 1944;55: 66–69. doi:10.3181/00379727-55-14461

298. Lewis K. Platforms for antibiotic discovery. Nat Rev Drug Discov. 2013;12: 371–387. doi:10.1038/nrd3975

299. Cozzoli D. Scott H. Podolsky, The Antibiotic Era: Reform, Resistance, and the Pursuit of a Rational Therapeutics. Baltimore: Johns Hopkins University Press, 2014. Pp. 328. ISBN 978-1-4214-1593-2. $34.95 (hardback). The British Journal for the History of Science. 2016. pp. 317–318. doi:10.1017/s0007087416000583

300. Klein EY, Van Boeckel TP, Martinez EM, Pant S, Gandra S, Levin SA, et al. Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. Proc Natl Acad Sci U S A. 2018;115: E3463–E3470. doi:10.1073/pnas.1717295115

301. Fridkin SK, Edwards JR, Courval JM, Hill H, Tenover FC, Lawton R, et al. The Effect of Vancomycin and Third-Generation Cephalosporins on Prevalence of Vancomycin-Resistant Enterococci in 126 U.S. Adult Intensive Care Units. Annals of Internal Medicine. 2001. p. 175. doi:10.7326/0003-4819-135-3-200108070-00009

302. Daneman N, Bronskill SE, Gruneir A, Newman AM, Fischer HD, Rochon PA, et al. Variability in Antibiotic Use Across Nursing Homes and the Risk of Antibiotic-Related Adverse Outcomes for Individual Residents. JAMA Intern Med. 2015;175: 1331–1339. doi:10.1001/jamainternmed.2015.2770

303. Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. BMJ. 2010;340: c2096. doi:10.1136/bmj.c2096

304. Steinke D, Davey P. Association between antibiotic resistance and community prescribing: a critical review of bias and confounding in published studies. Clin Infect Dis. 2001;33 Suppl 3: S193–205. doi:10.1086/321848

305. Goossens H, Ferech M, Vander Stichele R, Elseviers M, ESAC Project Group. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. Lancet. 2005;365: 579–587. doi:10.1016/S0140-6736(05)17907-0

306. Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance-the need for global solutions. Lancet Infect Dis. 2013;13: 1057–1098. doi:10.1016/S1473-3099(13)70318-9

307. Van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin SA, Robinson TP, et al. Global trends in antimicrobial use in food animals. Proc Natl Acad Sci U S A. 2015;112: 5649–5654. doi:10.1073/pnas.1503141112

308. Van Boeckel TP, Pires J, Silvester R, Zhao C, Song J, Criscuolo NG, et al. Global trends in antimicrobial resistance in animals in low- and middle-income countries. Science. 2019;365. doi:10.1126/science.aaw1944

309. Van Boeckel TP, Glennon EE, Chen D, Gilbert M, Robinson TP, Grenfell BT, et al. Reducing antimicrobial use in food animals. Science. 2017;357: 1350–1352. doi:10.1126/science.aao1495

310. The Lancet. Zoonoses: beyond the human-animal-environment interface. Lancet. 2020;396: 1. doi:10.1016/S0140-6736(20)31486-0

311. Vittecoq M, Godreuil S, Prugnolle F, Durand P, Brazier L, Renaud N, et al. Antimicrobial resistance in wildlife. McCallum H, editor. J Appl Ecol. 2016;53: 519–529. doi:10.1111/1365-2664.12596

312. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451: 990–993. doi:10.1038/nature06536

313. ECDC/EMEA Joint Technical Report: The bacterial challenge: time to react. ECDC–EMEA; 2009 Sep. Available: https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/0909_TER_The_Bacterial_Challenge_Time_to_React.pdf

314. Antibiotic resistance threats in the United States, 2019. Centers for Disease Control and Prevention; 2019. doi:10.15620/cdc:82532

315. Aminov RI. A brief history of the antibiotic era: lessons learned and challenges for the future. Front Microbiol. 2010;1: 134. doi:10.3389/fmicb.2010.00134

316. Cars O. Securing access to effective antibiotics for current and future generations. Whose responsibility? Ups J Med Sci. 2014;119: 209–214. doi:10.3109/03009734.2014.912700

317. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotic. World Health Organization; 2017 Feb. Available: https://www.who.int/medicines/publications/WHO-PPL-Short_Summary_25Feb-ET_NM_WHO.pdf?ua=1

318. De Oliveira DMP, Forde BM, Kidd TJ, Harris PNA, Schembri MA, Beatson SA, et al. Antimicrobial Resistance in ESKAPE Pathogens. Clin Microbiol Rev. 2020;33. doi:10.1128/CMR.00181-19

319. Theuretzbacher U, Outterson K, Engel A, Karlén A. The global preclinical antibacterial pipeline. Nat Rev Microbiol. 2020;18: 275–285. doi:10.1038/s41579-019-0288-0

320. Global action plan on antimicrobial resistance. World Health Organization; 2015. Available: https://apps.who.int/iris/bitstream/handle/10665/193736/9789241509763_eng.pdf?sequence=1

321. McEwen SA, Collignon PJ. Antimicrobial Resistance: a One Health Perspective. Microbiol Spectr. 2018;6. doi:10.1128/microbiolspec.ARBA-0009-2017

322. Amuasi JH, Lucas T, Horton R, Winkler AS. Reconnecting for our future: The Lancet One Health Commission. Lancet. 2020;395: 1469–1471. doi:10.1016/S0140-6736(20)31027-8

323. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet. 2019. doi:10.1038/s41576-019-0108-4

324. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front Microbiol. 2016;7: 173. doi:10.3389/fmicb.2016.00173

325. MacLean RC, San Millan A. The evolution of antibiotic resistance. Science. 2019;365: 1082–1083. doi:10.1126/science.aax3879

326. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev. 2018;31. doi:10.1128/CMR.00088-17

327. Bennett PM. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. Br J Pharmacol. 2008;153 Suppl 1: S347–57. doi:10.1038/sj.bjp.0707607

328. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, et al. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. Antimicrob Agents Chemother. 2016;60: 3767–3778. doi:10.1128/AAC.00464-16

329. Ambler RP, Coulson AFW, Frère J-M, Ghuysen J-M, Joris B, Forsman M, et al. A standard numbering scheme for the class A β-lactamases. Biochem J. 1991;276: 269–270. doi:10.1042/bj2760269

330. Bush K, Jacoby GA. Updated functional classification of beta-lactamases. Antimicrob Agents Chemother. 2010;54: 969–976. doi:10.1128/AAC.01009-09

331. Brandt C, Braun SD, Stein C, Slickers P, Ehricht R, Pletz MW, et al. In silico serine β-lactamases analysis reveals a huge potential resistome in environmental and pathogenic species. Sci Rep. 2017;7: 43232. doi:10.1038/srep43232

332. Nordmann P, Naas T, Poirel L. Global spread of carbapenemase-producing *Enterobacteriaceae*. Emerg Infect Dis. 2011;17: 1791. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3310682/

333. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect. 2017;23: 2–22. doi:10.1016/j.cmi.2016.11.012

334. Rossen JWA, Friedrich AW, Moran-Gilad J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin Microbiol Infect. 2018;424. doi:10.1016/j.cmi.2017.11.001

335. Angers-Loustau A, Petrillo M, Bengtsson-Palme J, Berendonk T, Blais B, Chan K-G, et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. F1000Res. 2018;7. doi:10.12688/f1000research.14509.2

336. Xu C, Wang D, Zhang X, Liu H, Zhu G, Wang T, et al. Mechanisms for Rapid Evolution of Carbapenem Resistance in a Clinical Isolate of *Pseudomonas aeruginosa*. Frontiers in Microbiology. 2020. doi:10.3389/fmicb.2020.01390

337. Wheeler NE, Sánchez-Busó L, Argimón S, Jeffrey B. Lean, mean, learning machines. Nature reviews. Microbiology. 2020. p. 266. doi:10.1038/s41579-020-0357-4

338. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. EBioMedicine. 2019;43: 356–369. doi:10.1016/j.ebiom.2019.04.016

339. Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO. A biochemically-interpretable machine learning classifier for microbial GWAS. Nat Commun. 2020;11: 2580. doi:10.1038/s41467-020-16310-9

340. Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab681

341. Pataki BÁ, Matamoros S, van der Putten BCL, Remondini D, Giampieri E, Aytan-Aktug D, et al. Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. Sci Rep. 2020;10: 15026. doi:10.1038/s41598-020-71693-5

342. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 2005;3: 711–721. doi:10.1038/nrmicro1234

343. Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of Plasmids. Microbiol Mol Biol Rev. 2010;74: 434–452. doi:10.1128/MMBR.00020-10

344. Carattoli A. Plasmids and the spread of resistance. Int J Med Microbiol. 2013;303: 298–304. doi:10.1016/j.ijmm.2013.02.001

345. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005;3: 722–732. doi:10.1038/nrmicro1235

346. Tazzyman SJ, Bonhoeffer S. Why There Are No Essential Genes on Plasmids. Mol Biol Evol. 2015;32: 3079–3088. doi:10.1093/molbev/msu293

347. Hamidian M, Holt KE, Pickard D, Hall RM. A small *Acinetobacter* plasmid carrying the tet39 tetracycline resistance determinant. J Antimicrob Chemother. 2016;71: 269–271. doi:10.1093/jac/dkv293

348. Guiney DG, Fang FC, Krause M, Libby S. Plasmid-mediated virulence genes in non-typhoid *Salmonella* serovars. FEMS Microbiol Lett. 1994;124: 1–9. doi:10.1111/j.1574-6968.1994.tb07253.x

349. Carattoli A, Villa L, Fortini D, García-Fernández A. Contemporary IncI1 plasmids involved in the transmission and spread of antimicrobial resistance in *Enterobacteriaceae*. Plasmid. 2018. doi:10.1016/j.plasmid.2018.12.001

350. Dolejska M, Papagiannitsis CC. Plasmid-mediated resistance is going wild. Plasmid. 2018;99: 99–111. doi:10.1016/j.plasmid.2018.09.010

351. Wick RR, Judd LM, Wyres KL, Holt KE. Recovery of small plasmid sequences via Oxford Nanopore sequencing. Microb Genom. 2021;7. doi:10.1099/mgen.0.000631

352. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom. 2017;3: e000132. doi:10.1099/mgen.0.000132

353. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58: 3895–3903. doi:10.1128/AAC.02412-14

354. Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, et al. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. Microb Genom. 2018;4. doi:10.1099/mgen.0.000211

355. Roosaare M, Puustusmaa M, Möls M, Vaher M, Remm M. PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ. 2018;6: e4588. doi:10.7717/peerj.4588

356. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics. 2010;26: 2051–2052. doi:10.1093/bioinformatics/btq299

357. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. 2018;46: e35–e35. doi:10.1093/nar/gkx1321

358. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. Mlplasmids: a User-Friendly Tool To Predict Plasmid- and Chromosome-Derived Sequences for Single Species. Microbial Genomics. 2018; 1–15. doi:10.1099/mgen.0.000224

359. Pellow D, Mizrahi I, Shamir R. PlasClass improves plasmid sequence classification. PLoS Comput Biol. 2020;16: e1007781. doi:10.1371/journal.pcbi.1007781

360. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics. 2016;32: 3380–3387. doi:10.1093/bioinformatics/btw493

361. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, et al. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics. 2016;33: btw651. doi:10.1093/bioinformatics/btw651

362. Vielva L, de Toro M, Lanza VF, de la Cruz F. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics. 2017;33: 3796–3798. doi:10.1093/bioinformatics/btx462

363. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, et al. gplas: a comprehensive tool for plasmid analysis using short-read graphs. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa233

364. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microbial Genomics. 2017; 1–8. doi:10.1099/mgen.0.000128

365. Laczny CC, Galata V, Plum A, Posch AE, Keller A. Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. Brief Bioinform. 2019;20: 857–865. doi:10.1093/bib/bbx162

366. Hilpert C, Bricheux G, Debroas D. Reconstruction of plasmids by shotgun sequencing from environmental DNA: which bioinformatic workflow? Brief Bioinform. 2020. doi:10.1093/bib/bbaa059

367. GenBank and WGS Statistics. [cited 26 Jan 2021]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/

368. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, et al. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res. 2016;44: D73–80. doi:10.1093/nar/gkv1226

369. Ruffier M, Kähäri A, Komorowska M, Keenan S, Laird M, Longden I, et al. Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. Database . 2017;2017. doi:10.1093/database/bax020

370. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

371. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013;10: e1001387. doi:10.1371/journal.pmed.1001387

372. Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics. 2017;18: 474. doi:10.1186/s12859-017-1911-6

373. Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Systematic analysis of key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria. 2020. p. 2020.09.24.310821. doi:10.1101/2020.09.24.310821

374. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132. doi:10.1186/s13059-016-0997-x

375. ANI calculator. In: Kostas lab [Internet]. [cited 11 Nov 2020]. Available: http://enve-omics.ce.gatech.edu/ani/

376. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol. 2016;66: 1100–1103. doi:10.1099/ijsem.0.000760

377. Zhao X. BinDash, software for fast genome distance estimation on a typical personal laptop. Bioinformatics. 2019;35: 671–673. doi:10.1093/bioinformatics/bty651

378. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci U S A. 2012;109: 4550–4555. doi:10.1073/pnas.1113219109

379. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, et al. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. PLoS One. 2013;8: e61319. doi:10.1371/journal.pone.0061319

380. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open. 2012;2. doi:10.1136/bmjopen-2012-001124

381. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13: 137–146. doi:10.1016/S1473-3099(12)70277-3

382. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir Med. 2014;2: 285–292. doi:10.1016/S2213-2600(14)70027-X

383. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011;477: 462–465. doi:10.1038/nature10392

384. Field D, Wilson G, van der Gast C. How do we compare hundreds of bacterial genomes? Curr Opin Microbiol. 2006;9: 499–504. doi:10.1016/j.mib.2006.08.008

385. Moore GE. Cramming more components onto integrated circuits. Electronics, 38 (8). April; 1965.

386. Walter C. Kryder's Law. Scientific American. 1 Aug 2005. doi:10.1038/scientificamerican0805-32

387. The 100,000 Genomes Project. 21 Jul 2014 [cited 9 Dec 2021]. Available: https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/

388. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449: 804–810. doi:10.1038/nature06244

389. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551: 457–463. doi:10.1038/nature24621

390. Mell P, Grance T. The NIST Definition of Cloud Computing. National Institute of Standards and Technology; 2011 Sep. Report No.: NIST Special Publication (SP) 800-145. doi:10.6028/NIST.SP.800-145

391. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, et al. A view of cloud computing. Commun ACM. 2010;53: 50–58. doi:10.1145/1721654.1721672

392. Singh B, Kaur J, Singh K. Microbial remediation of explosive waste. Crit Rev Microbiol. 2012;38: 152–167. doi:10.3109/1040841X.2011.640979

393. Wackett LP. Microbial-based motor fuels: science and technology. Microb Biotechnol. 2008;1: 211–225. doi:10.1111/j.1751-7915.2007.00020.x

394. Schempp FM, Drummond L, Buchhaupt M, Schrader J. Microbial Cell Factories for the Production of Terpenoid Flavor and Fragrance Compounds. J Agric Food Chem. 2018;66: 2247–2258. doi:10.1021/acs.jafc.7b00473

395. Ison J, Rapacki K, Ménager H, Kalaš M, Rydza E, Chmura P, et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. 2016;44: D38–47. doi:10.1093/nar/gkv1116

396. Felter W, Ferreira A, Rajamony R, Rubio J. An updated performance comparison of virtual machines and Linux containers. 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). 2015. pp. 171–172. doi:10.1109/ISPASS.2015.7095802

397. Chae M, Lee H, Lee K. A performance comparison of linux containers and virtual machines using Docker and KVM. Cluster Comput. 2019;22: 1765–1775. doi:10.1007/s10586-017-1511-2

398. Lawlor B, Sleator RD. The democratization of bioinformatics: A software engineering perspective. Gigascience. 2020;9. doi:10.1093/gigascience/giaa063

399. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3: e1273. doi:10.7717/peerj.1273

400. Boettiger C. An introduction to Docker for reproducible research. Oper Syst Rev. 2015;49: 71–79. doi:10.1145/2723872.2723882

401. Open Grid Scheduler. [cited 29 Aug 2020]. Available: http://gridscheduler.sourceforge.net

402. Slurm Workload Manager - Documentation. [cited 29 Aug 2020]. Available: https://slurm.schedmd.com/documentation.html

403. OpenPBS Open Source Project. [cited 29 Aug 2020]. Available: https://www.openpbs.org

404. Kubernetes. [cited 29 Aug 2020]. Available: https://kubernetes.io

405. AWS Batch. [cited 29 Aug 2020]. Available: https://aws.amazon.com/de/batch

406. Cloud Life Sciences. [cited 29 Aug 2020]. Available: https://cloud.google.com/life-sciences

407. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522. doi:10.1093/bioinformatics/bts480

408. Marx V. Pipeline plunk. In: Springer Nature Protocols and Methods Community [Internet]. Springer Nature; 23 Jun 2020 [cited 9 Jul 2020]. Available: http://protocolsmethods.springernature.com/users/59087-vivien-marx/posts/pipeline-plunk

409. Belmann P, Fischer B, Krüger J, Procházka M, Rasche H, Prinz M, et al. de.NBI Cloud federation through ELIXIR AAI. F1000Res. 2019;8: 842. doi:10.12688/f1000research.19013.1

410. Gargis AS, Kalman L, Lubin IM. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol. 2016;54: 2857–2865. doi:10.1128/JCM.00949-16

411. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol. 2021;19: e3001421. doi:10.1371/journal.pbio.3001421

412. Sinville R, Soper SA. High resolution DNA separations using microchip electrophoresis. J Sep Sci. 2007;30: 1714–1728. doi:10.1002/jssc.200700150

413. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351. doi:10.1038/nrg.2016.49

414. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. Semin Thromb Hemost. 2019;45: 661–673. doi:10.1055/s-0039-1688446

415. Applied Biosystems Genetic Analysis Systems - DE. [cited 27 Jan 2021]. Available: https://www.thermofisher.com/de/de/home/life-science/sequencing/sanger-sequencing/sanger-sequencing-technology-accessories.html

416. GenBank and WGS Statistics. [cited 10 Nov 2020]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/

417. DNA Sequencing Costs: Data. [cited 26 Jan 2021]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

418. Thomsen MCF, Ahrenfeldt J, Cisneros JLB, Jurtz V, Larsen MV, Hasman H, et al. A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. PLoS One. 2016;11: e0157718. doi:10.1371/journal.pone.0157718

419. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46: W537–W544. doi:10.1093/nar/gky379

420. Müller R, Chauve C. HyAsP, a greedy tool for plasmids identification. Bioinformatics. 2019;35: 4436–4439. doi:10.1093/bioinformatics/btz413

421. Xavier BB, Mysara M, Bolzan M, Ribeiro-Gonçalves B, Alako BTF, Harrison P, et al. BacPipe: A Rapid, User-Friendly Whole-Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology. iScience. 2020;23: 100769. doi:10.1016/j.isci.2019.100769

422. Quijada NM, Rodríguez-Lázaro D, Hernández M. TORMES: an automated pipeline for whole bacterial genome analysis. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz220

423. Seemann T. nullarbor. Github; Available: https://github.com/tseemann/nullarbor

424. Sserwadda I, Mboowa G. rMAP: the Rapid Microbial Analysis Pipeline for ESKAPE bacterial group whole-genome sequence data. Microb Genom. 2021;7. doi:10.1099/mgen.0.000583

425. Pavlovikj N, Gomes-Neto JC, Deogun JS, Benson AK. ProkEvo: an automated, reproducible, and scalable framework for high-throughput bacterial population genomics analyses. PeerJ. 2021;9: e11376. doi:10.7717/peerj.11376

426. Petit RA 3rd, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 2020;5. doi:10.1128/mSystems.00190-20

427. Falgenhauer J. 2020.

428. Deutsches Netzwerk Für Bioinformatik-Infrastruktur. Von der Datenanalyse zum Verstehen komplexer biologischer Systeme - Highlights aus dem Deutschen Netzwerk für Bioinformatik-Infrastruktur. 2020 Jan. Available: https://www.denbi.de/news-archive/831-new-highlight-brochure-in-german-available

429. Mulani MS, Kamble EE, Kumkar SN, Tawre MS, Pardesi KR. Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. Front Microbiol. 2019;10: 539. doi:10.3389/fmicb.2019.00539

430. Ayeni FA, Falgenhauer J, Schmiedel J, Schwengers O, Chakraborty T, Falgenhauer L. Detection of blaCTX-M-27-encoding *Escherichia coli* ST206 in Nigerian poultry stocks. J Antimicrob Chemother. 2020. doi:10.1093/jac/dkaa293

431. Falgenhauer L, Imirzalioglu C, Oppong K, Akenten CW, Hogan B, Krumkamp R, et al. Detection and Characterization of ESBL-Producing *Escherichia* coli From Humans and Poultry in Ghana. Front Microbiol. 2018;9: 3358. doi:10.3389/fmicb.2018.03358

432. Falgenhauer L, Nordmann P, Imirzalioglu C, Yao Y, Falgenhauer J, Hauri AM, et al. Cross-border emergence of clonal lineages of ST38 *Escherichia coli* producing the OXA-48-like carbapenemase OXA-244 in Germany and Switzerland. Int J Antimicrob Agents. 2020; 106157. doi:10.1016/j.ijantimicag.2020.106157

433. Yin Y, Doijad S, Wang W, Lian K, Pan X, Koryciński I, et al. Genetic Diversity of *Listeria monocytogenes* Isolates from Invasive Listeriosis in China. Foodborne Pathog Dis. 2020;17: 215–227. doi:10.1089/fpd.2019.2693

434. Perniss A, Schmidt N, Gurtner C, Dietert K, Schwengers O, Weigel M, et al. *Bordetella pseudohinzii* targets cilia and impairs tracheal cilia-driven transport in naturally acquired infection in mice. Sci Rep. 2018;8: 5681. doi:10.1038/s41598-018-23830-4

435. Pappesch R, Warnke P, Mikkat S, Normann J, Wisniewska-Kucper A, Huschka F, et al. The Regulatory Small RNA MarS Supports Virulence of *Streptococcus pyogenes*. Sci Rep. 2017;7: 12241. doi:10.1038/s41598-017-12507-z

436. Falgenhauer J, Imirzalioglu C, Falgenhauer L, Yao Y, Hauri AM, Erath B, et al. Whole-Genome Sequences of Clinical *Enterobacter bugandensis* Isolates from Germany. Microbiol Resour Announc. 2019;8. doi:10.1128/MRA.00465-19

437. Canchaya C, Fournous G, Brüssow H. The impact of prophages on bacterial chromosomes. Mol Microbiol. 2004;53: 9–18. doi:10.1111/j.1365-2958.2004.04113.x

438. Pan Y, Fang Y, Feng Y, Lyu N, Chen L, Li J, et al. Discovery of mcr-3.1 gene carried by a prophage located in a conjugative IncA/C2 plasmid from a *Salmonella choleraesuis* clinical isolate. J Infect. 2020. doi:10.1016/j.jinf.2020.09.036

439. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15: R46. doi:10.1186/gb-2014-15-3-r46

440. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. Microbial Genomics. 2020;95. doi:10.1099/mgen.0.000398

441. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, et al. ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. PLoS Comput Biol. 2020;16: e1007134. doi:10.1371/journal.pcbi.1007134

442. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35: 316–319. doi:10.1038/nbt.3820

443. Andreopoulos WB, Geller AM, Lucke M, Balewski J, Clum A, Ivanova NN, et al. Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. Nucleic Acids Res. 2021. doi:10.1093/nar/gkab1115

444. Schwengers O, Hain T, Chakraborty T, Goesmann A. ReferenceSeeker: rapid determination of appropriate reference genomes. JOSS. 2020;5: 1994. doi:10.21105/joss.01994

445. Schwengers O. ReferenceSeeker Database. Zenodo; 2020. doi:10.5281/ZENODO.3992357

7 Publications

- **ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates.**
  Oliver Schwengers, Andreas Hoek, Moritz Fritzenwanker, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).
  PLoS Computational Biology, DOI: 10.1371/journal.pcbi.1007134

- **Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein-sequence-based replicon distribution scores.**
  Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).
  Microbial Genomics, DOI: 10.1099/mgen.0.000398

- **ReferenceSeeker: rapid determination of appropriate reference genomes.**
  Oliver Schwengers, Torsten Hain, Trinad Chakraborty & Alexander Goesmann (2020).
  Journal of Open Source Software, DOI: 10.21105/joss.01994

7.1 ASA³P: an automatic and scalable pipeline for the assembly, annotation and higher level analysis of closely related bacterial isolates.

Oliver Schwengers[1,2,3 *], Andreas Hoek[1], Moritz Fritzenwanker[2,3],
Linda Falgenhauer[2,3], Torsten Hain[2,3], Trinad Chakraborty[2,3,‡],
Alexander Goesmann[1,3,‡]

Affiliations	[1] Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany [2] Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, Germany [3] German Center for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Giessen, Germany * Corresponding author ‡ These authors contributed equally to this work.

Journal	PLoS Computational Biology
Date	2020
Volume	16(3)
DOI	10.1371/journal.pcbi.1007134

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates

Oliver Schwengers[1,2,3]*, **Andreas Hoek**[1], **Moritz Fritzenwanker**[2,3], **Linda Falgenhauer**[2,3], **Torsten Hain**[2,3], **Trinad Chakraborty**[2,3☯], **Alexander Goesmann**[1,3☯]

1 Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany, **2** Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, Germany, **3** German Center for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Giessen, Germany

☯ These authors contributed equally to this work.
* oliver.schwengers@computational.bio.uni-giessen.de

OPEN ACCESS

Citation: Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, et al. (2020) ASA^{3}P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. PLoS Comput Biol 16(3): e1007134. https://doi.org/10.1371/journal.pcbi.1007134

Editor: Mihaela Pertea, Johns Hopkins University, UNITED STATES

Received: May 24, 2019

Accepted: December 3, 2019

Published: March 5, 2020

Data Availability Statement: All source code is accessible at GitHub (https://github.com/oschwengers/asap). The software bundle, manual, exemplary data sets, etc. are accessible via Download at Zenodo.org. DOIs and download URLs are provided in the GitHub repository readme as well as our institutional software page (https://www.uni-giessen.de/fbz/fb08/Inst/bioinformatik/software/asap). Genomes from the exemplary data sets are publicly stored in the SRA database;

Abstract

Whole genome sequencing of bacteria has become daily routine in many fields. Advances in DNA sequencing technologies and continuously dropping costs have resulted in a tremendous increase in the amounts of available sequence data. However, comprehensive in-depth analysis of the resulting data remains an arduous and time-consuming task. In order to keep pace with these promising but challenging developments and to transform raw data into valuable information, standardized analyses and scalable software tools are needed. Here, we introduce ASA^{3}P, a fully automatic, locally executable and scalable assembly, annotation and analysis pipeline for bacterial genomes. The pipeline automatically executes necessary data processing steps, *i.e.* quality clipping and assembly of raw sequencing reads, scaffolding of contigs and annotation of the resulting genome sequences. Furthermore, ASA^{3}P conducts comprehensive genome characterizations and analyses, *e.g.* taxonomic classification, detection of antibiotic resistance genes and identification of virulence factors. All results are presented via an HTML5 user interface providing aggregated information, interactive visualizations and access to intermediate results in standard bioinformatics file formats. We distribute ASA^{3}P in two versions: a locally executable Docker container for small-to-medium-scale projects and an OpenStack based cloud computing version able to automatically create and manage self-scaling compute clusters. Thus, automatic and standardized analysis of hundreds of bacterial genomes becomes feasible within hours. The software and further information is available at: asap.computational.bio.

This is a *PLOS Computational Biology* Software paper.

Introduction

In 1977 DNA sequencing was introduced to the scientific community by Frederick Sanger [1]. Since then, DNA sequencing has come a long way from dideoxy chain termination over high

accession IDs are provided in the supporting information.

Funding: This work was supported by the German Center of Infection Research (DZIF) (DZIF grant 8000 701–3 [HZI], TI06.001, 8032808811, and 8032808820 to TC); the German Network for Bioinformatics Infrastructure (de.NBI) (BMBF grant FKZ 031A533B to AG); and the German Research Foundation (DFG) (SFB-TR84 project A04 [TRR84/3 2018] to TC, KFO309 Z1 [GO 2037/5-1] to AG, SFB-TR84 project B08 [TRR84/3 2018] to TH, SFB1021 Z02 [SFB 1021/2 2017] to TH, KFO309 Z1 [HA 5225/1-1] to TH). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

throughput sequencing of millions of short DNA fragments and finally to real-time sequencing of single DNA molecules [2,3]. Latter technologies of so-called next generation sequencing (NGS) and third generation sequencing have caused a massive reduction of time and costs, and thus, led to an explosion of publicly available genomes. In 1995, the first bacterial genomes of *M. genitalium* and *H. influenzae* were published [4,5]. Today, the NCBI RefSeq database release 93 alone contains 54,854 genomes of distinct bacterial organisms [6]. Due to the maturation of NGS technologies, the laborious task of bacterial whole genome sequencing (WGS) has transformed into plain routine [7] and nowadays, has become feasible within hours [8].

As the sequencing process is not a limiting factor anymore, focus has shifted towards deeper analyses of single genomes and also large cohorts of *e.g.* clinical isolates in a comparative way to unravel the plethora of genetic mechanisms driving diversity and genetic landscape of bacterial populations [9]. Comprehensively characterizing bacterial organisms has become a desirable and necessary task in many fields of application including environmental- and medical microbiology [10]. The recent worldwide surge of multi-resistant microorganisms has led to the realization, that without the implementation of adequate measures in 2050 up to 10 million people could die each year due to infections with antimicrobial resistant bacteria alone [11]. Thus, sequencing and timely characterization of large numbers of bacterial genomes is a key element for successful outbreak detection, proper surveillance of emerging pathogens and monitoring the spread of antibiotic resistance genes [12]. Comparative analysis could lead to the identification of novel therapeutic drug targets to prevent the spread of pathogenic and antibiotic-resistant bacteria [13–16].

Another very promising and important field of application for microbial genome sequencing is modern biotechnology. Due to deeper knowledge of the underlying genomic mechanisms, genetic engineering of genes and entire bacterial genomes has become an indispensable tool to transform them into living chemical factories with vast applications, as for instance, production of complex chemicals [17], synthesis of valuable drugs [18–20] and biofuels [21], decontamination and degradation of toxins and wastes [22,23] as well as corrosion protection [24].

Now, that the technological barriers of WGS have fallen, genomics finally transformed into Big Data science [25] inducing new issues and challenges [26]. To keep pace with these developments, we believe that continued efforts are required in terms of the following issues:

a) Automation: Repeated manual analyses are time consuming and error prone. Following the well-known “don’t repeat yourself” mantra and the pareto principle, scientists should be able to concentrate on interesting and promising aspects of data analysis instead of ever repeating data processing tasks.

b) Standard operating procedures (SOPs): In a world of high-throughput data creation and complex combinations of bioinformatic tools SOPs are indispensable to increase and maintain both reproducibility and comparability [27].

c) Scalability: To keep pace with the available data, bioinformatics software needs to take advantage of modern computing technologies, *e.g.* multi-threading and cloud computing.

Addressing these issues, several major platforms for the automatic annotation and analysis of prokaryotic genomes have evolved in recent years as for example the NCBI Prokaryotic Genome Annotation Pipeline [6], RAST [28] and PATRIC [29]. All three provide sophisticated genome analysis and annotation pipelines and pose a de-facto community standard in terms of annotation quality. In addition, several offline tools, *e.g.* Prokka [30], have been published in order to address major drawbacks of the aforementioned online tools, *i.e.* they are not executable on local computers or in on-premises cloud computing environments. However, comprehensive analysis of bacterial WGS data is not limited to the process of annotation alone but also requires sequencing technology-dependent pre-processing of raw data as well as subsequent characterization steps. As analysis of bacterial isolates and cohorts will be a standard method in many fields of application in the near future, demand for sophisticated local assembly, annotation and higher-level analysis pipelines will rise constantly. Furthermore, we believe that the utilization of portable devices for DNA sequencing will shift analysis from central software installations to either decentral offline tools or scalable cloud solutions. To the authors' best knowledge, there is currently no published bioinformatics software tool successfully addressing all aforementioned issues. In order to overcome this bottleneck, we introduce ASA^{3}P, an automatic and scalable software pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates.

Design and implementation

ASA^{3}P is implemented as a modular command line tool in Groovy (http://groovy-lang.org), a dynamic scripting language for the Java virtual machine. In order to achieve acceptable to best possible results over a broad range of bacterial genera, sequencing technologies and sequencing depths, ASA^{3}P incorporates and takes advantage of published and well performing bioinformatics tools wherever available and applicable in terms of lean and scalable implementation. As the pipeline is also intended to be used as a preprocessing tool for more specialized analyses, it provides no user-adjustable parameters by design and thus facilitates the implementation of robust SOPs. Hence, each utilized tool is parameterized according to community best practices and knowledge (**S1 Table**).

Workflow, tools and databases

Depending on the sequencing technology used to generate the data, ASA^{3}P automatically chooses appropriate tools and parameters. An explanation on which tool was chosen for each task is given in **S2 Table**. Semantically, the pipeline's workflow is divided into four stages (**Fig 1**). In the first mandatory stage A (**Fig 1A**), provided input data are processed, resulting in annotated genomes. Therefore, raw sequencing reads are quality controlled and clipped via FastQC (https://github.com/s-andrews/FastQC), FastQ Screen (https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen), Trimmomatic [31] and Filtlong (https://github.com/rrwick/Filtlong). Filtered reads are then assembled via SPAdes [32] for Illumina reads, HGAP 4 [33] for Pacific Bioscience (PacBio) reads and Unicycler [34] for Oxford Nanopore Technology (ONT) reads, respectively. Hybrid assemblies of Illumina and ONT reads are conducted via Unicycler, as well. Before annotating assembled genomes with Prokka [30], contigs are rearranged and ordered via the multi-reference scaffolder MeDuSa [35]. For the annotation of subsequent pseudogenomes ASA^{3}P uses custom genus-specific databases based on binned RefSeq genomes [6] as well as specialized protein databases, *i.e.* CARD [36] and VFDB [37]. In order to integrate public or externally analyzed genomes, ASA^{3}P is able to incorporate different types of pre-processed data, *e.g.* contigs, scaffolds and annotated genomes.

In an optional second stage B (**Fig 1B**), all assembled and annotated genomes are extensively characterized. A taxonomic classification is conducted comprising three distinct methods, *i.e.* a kmer profile search, a 16S sequence homology search and a computation of an average nucleotide identity (ANI) [38] against user provided reference genomes. For the kmer profile search, the software takes advantage of the Kraken package [39] and a custom reference genome database based on RefSeq [6]. The 16S based classification is implemented using BLAST+ [40] and the SILVA [41] database. Calculation of ANI values is implemented in Groovy using nucmer within the MUMmer package [42]. A subspecies level multi locus sequence typing (MLST) analysis is implemented in Groovy using BLAST+ [40] and the PubMLST.org [43] database. Detection of antibiotic resistances (ABRs) is conducted via RGI

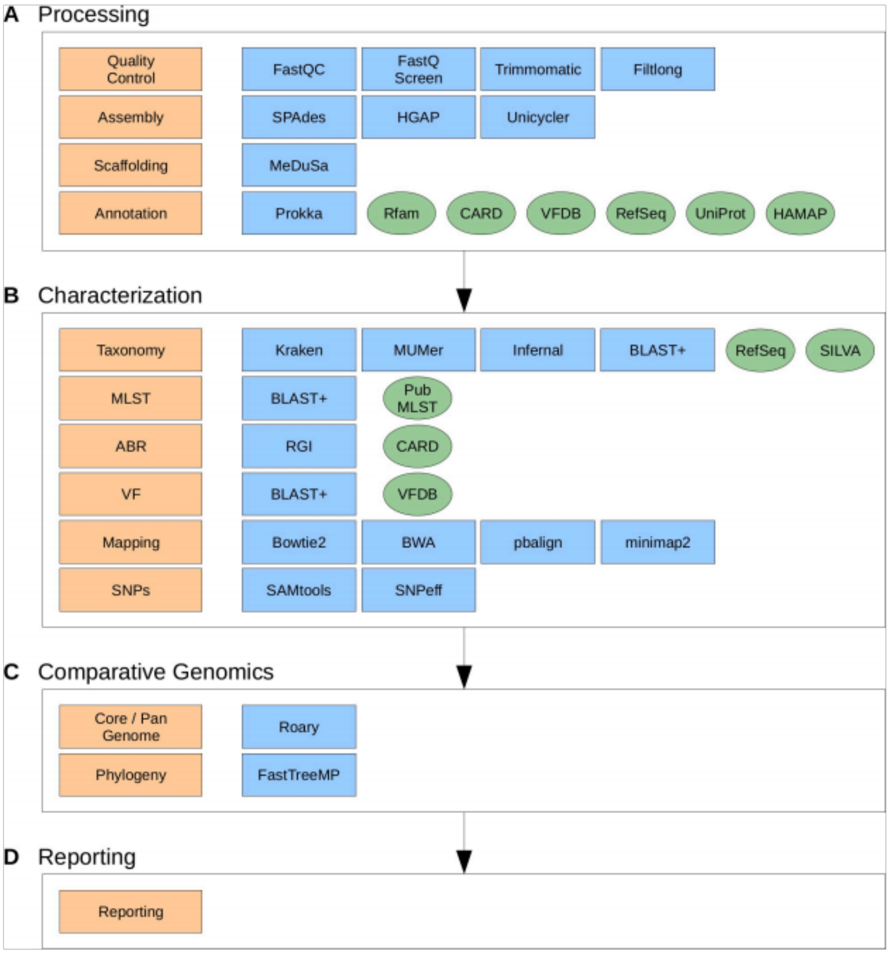


Fig 1. The ASA^{3}P workflow and incorporated third party software tools and databases. The ASA^{3}P workflow is organized in four stages (large white boxes, A-D) comprising per-isolate processing and characterization, comparative analysis and reporting steps (orange boxes). The processing stage A is mandatory whereas stage B and C are optional and can be skipped by the user. Each step takes advantage of selected third-party software tools (blue boxes) and/or databases (green ovals) depending on the type of provided input data at hand.

https://doi.org/10.1371/journal.pcbi.1007134.g001

and the CARD [36] database. A detection of virulence factors (VFs) is implemented via BLAST+ [40] and VFDB [37]. Quality clipped reads get mapped onto user provided reference genomes via Bowtie2 [44] for Illumina, pbalign (https://github.com/PacificBiosciences/pbalign) for PacBio and Minimap2 [45] for ONT sequence reads, respectively. Based on these read mappings, the pipeline calls, filters and annotates SNPs via SAMtools [46] and SnpEff [47] and finally computes consensus sequences for each isolate. In order to maximize parallel execution and thus reducing overall runtime, stage A and B are technically implemented as a single step.

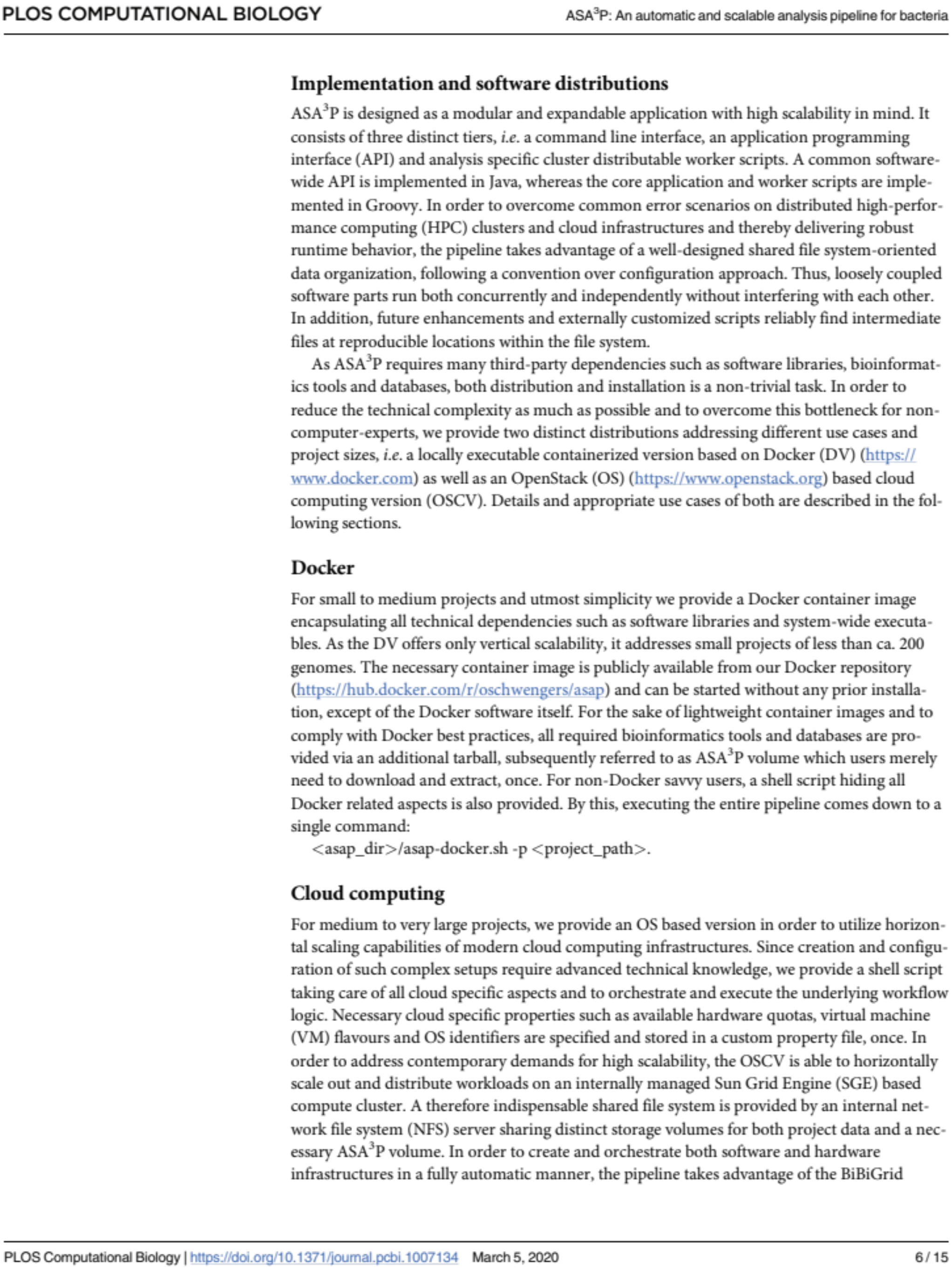
A third optional comparative stage C (Fig 1C) is triggered as soon as stages A and B are completed, *i.e.* all genomes are processed and characterized. Utilizing aforementioned consensus sequences, ASA^{3}P computes a phylogenetic approximately maximum-likelihood tree via FastTreeMP [48]. This is complemented by the calculation of a core, accessory and pan-genome as well as the detection of isolate genes conducted via Roary [49].

In a final stage (Fig 1D), the pipeline aggregates analysis results and data files and finally provides a graphical user interface (GUI), *i.e.* responsive HTML5 documents comprising detailed information via interactive widgets and visualizations. Therefore, ASA^{3}P takes advantage of modern web frameworks, *e.g.* Bootstrap (https://getbootstrap.com) and jQuery (https://jquery.com) as well as adequate JavaScript visualization libraries, *e.g.* Google Charts (https://developers.google.com/chart), D3 (https://d3js.org) and C3 (http://c3js.org).

User input and output

Each set of bacterial isolates to be analyzed within a single execution is considered as a self-contained analysis of bacterial cohorts and is subsequently referred to as an ASA^{3}P project. As ASA^{3}P was developed in order to analyze cohorts of closely related isolates, *e.g.* a clonal outbreak, the pipeline expects all genomes within a project to belong to at least the same genus, although a common species is most favourable. For each project, the pipeline expects a distinct directory comprising a configuration spreadsheet containing necessary project information and a subdirectory containing all input data files. Such a directory is subsequently referred to as project directory. In order to ease provisioning of necessary information, we provide a configuration spreadsheet template comprising two sheets (**S1 and S2 Figs**). The first sheet contains project meta information such as project names and descriptions as well as contact information on project maintainers and provided reference genomes. The second sheet stores information on each isolate comprising a unique identifier as well as data input type and related files. ASA^{3}P is currently able to process input data in the following standard file formats: Illumina paired-end and single-end reads as compressed FastQ files, PacBio RSII and Sequel reads provided either as single unmapped bam files or via triples of bax.h5 files, demultiplexed ONT reads as compressed FastQ files, pre-assembled contigs or pseudogenomes as Fasta files and pre-annotated genomes as Genbank, EMBL or GFF files. In the latter case, corresponding genome sequences can either be included in the GFF file or provided via separate Fasta files.

As ASA^{3}P is also intended to be used as an automatic preprocessing tool providing as much reliable information as possible, the results are stored in a standardized manner within project directories comprising quality clipped reads, assemblies, ordered and scaffolded contigs, annotated genomes, mapped reads, detected SNPs as well as ABRs and VFs. In detail, all result files are stored in distinct subdirectories for each analysis by the pipeline and for certain analyses further subdirectories are created therein for each genome (**S3 Fig**). Aggregated information is stored in a standardized but flexible document structure as JSON files. Text and binary result files are stored in standard bioinformatics file formats, *i.e.* FastQ, Fasta, BAM, VCF and Newick. Providing results in such a machine-readable manner, ASA^{3}P outputs can be further exploited by manual or automatic downstream analyses since customized scripts with a more targeted focus can easily access necessary data. In addition, ASA^{3}P creates user-friendly HTML5 reports providing both prepared summaries as well as detailed information via sophisticated interactive visualizations.



Implementation and software distributions

ASA^{3}P is designed as a modular and expandable application with high scalability in mind. It consists of three distinct tiers, *i.e.* a command line interface, an application programming interface (API) and analysis specific cluster distributable worker scripts. A common software-wide API is implemented in Java, whereas the core application and worker scripts are implemented in Groovy. In order to overcome common error scenarios on distributed high-performance computing (HPC) clusters and cloud infrastructures and thereby delivering robust runtime behavior, the pipeline takes advantage of a well-designed shared file system-oriented data organization, following a convention over configuration approach. Thus, loosely coupled software parts run both concurrently and independently without interfering with each other. In addition, future enhancements and externally customized scripts reliably find intermediate files at reproducible locations within the file system.

As ASA^{3}P requires many third-party dependencies such as software libraries, bioinformatics tools and databases, both distribution and installation is a non-trivial task. In order to reduce the technical complexity as much as possible and to overcome this bottleneck for non-computer-experts, we provide two distinct distributions addressing different use cases and project sizes, *i.e.* a locally executable containerized version based on Docker (DV) (https://www.docker.com) as well as an OpenStack (OS) (https://www.openstack.org) based cloud computing version (OSCV). Details and appropriate use cases of both are described in the following sections.

Docker

For small to medium projects and utmost simplicity we provide a Docker container image encapsulating all technical dependencies such as software libraries and system-wide executables. As the DV offers only vertical scalability, it addresses small projects of less than ca. 200 genomes. The necessary container image is publicly available from our Docker repository (https://hub.docker.com/r/oschwengers/asap) and can be started without any prior installation, except of the Docker software itself. For the sake of lightweight container images and to comply with Docker best practices, all required bioinformatics tools and databases are provided via an additional tarball, subsequently referred to as ASA^{3}P volume which users merely need to download and extract, once. For non-Docker savvy users, a shell script hiding all Docker related aspects is also provided. By this, executing the entire pipeline comes down to a single command:

<asap_dir>/asap-docker.sh -p <project_path>.

Cloud computing

For medium to very large projects, we provide an OS based version in order to utilize horizontal scaling capabilities of modern cloud computing infrastructures. Since creation and configuration of such complex setups require advanced technical knowledge, we provide a shell script taking care of all cloud specific aspects and to orchestrate and execute the underlying workflow logic. Necessary cloud specific properties such as available hardware quotas, virtual machine (VM) flavours and OS identifiers are specified and stored in a custom property file, once. In order to address contemporary demands for high scalability, the OSCV is able to horizontally scale out and distribute workloads on an internally managed Sun Grid Engine (SGE) based compute cluster. A therefore indispensable shared file system is provided by an internal network file system (NFS) server sharing distinct storage volumes for both project data and a necessary ASA^{3}P volume. In order to create and orchestrate both software and hardware infrastructures in a fully automatic manner, the pipeline takes advantage of the BiBiGrid (https://github.com/BiBiServ/bibigrid/) framework. Hereby, ASA^{3}P is able to adjust the compute cluster size fitting the number of isolates within a project as well as available hardware quotas. Except of an initial VM acting as a gateway into an OS cloud project, the entire compute cluster infrastructure is automatically created, setup, managed and finally shut down by the software. Thus, ASA^{3}P can exploit vast hardware capacities and is portable to any OS compatible cloud. For further guidance, all prerequisite installation steps are covered in a detailed user manual.

Results

Analysis features

ASA^{3}P conducts a comprehensive set of pre-processing tasks and genome analyses. In order to delineate currently implemented analysis features, we created and analyzed a benchmark data set comprising 32 Illumina sequenced *Listeria monocytogenes* isolates randomly selected from SRA as well as four *Listeria monocytogenes* reference genomes from Genbank (S3 Table). All isolates were successfully assembled, annotated, deeply characterized and finally included in comparative analyses. Table 1 provides genome wise minimum and maximum values for key metrics covering results from workflow stages A and B. After conducting a quality control and adapter removal for all raw sequencing reads, a minimum of 393,300 and a maximum of 6,315,924 reads remained, respectively. Genome wise minimum and maximum mean phred scores were 34.7 and 37.2. Assembled genome sizes ranged between 2,818 kbp and 3,201 kbp with a minimum of 12 and a maximum of 108 contigs. Hereby, a maximum N50 of 1,568 kbp was achieved. After rearranging and ordering contigs to aforementioned reference genomes, assemblies were reduced to 2 to 10 scaffolds and 0 to 42 contigs per genome, thus increasing the minimum and maximum N50 to 658 kbp and 3,034 kbp, respectively. Pseudolinked genomes were subsequently annotated resulting in between 2,735 and 3,200 coding genes and between 95 and 144 non-coding genes.

After pre-processing, assembling and annotating all isolates, ASA^{3}P successfully conducted deep characterizations of all isolates, which were consistently classified to the species level via

Table 1. Common genome analysis key metrics for processing and characterization steps analyzing a benchmark dataset comprising 32 *Listeria monocytogenes* isolates. Minimum and maximum values for selected common genome analysis key metrics resulting from an automatic analysis conducted with ASA^{3}P of an exemplary benchmark dataset comprising 32 *Listeria monocytogenes* isolates. Metrics are given for quality control (QC), assembly, scaffolding and annotation processing steps as well as detection of antibiotic resistances and virulence factors characterization steps on a per-isolate level.

Analysis	Metric	Minimum	Maximum
QC	reads	393,300	6,315,924
QC	Mean read length	125.7 nt	228.5 nt
QC	mean Phred score	34.7	37.2
assembly	Genome size	2,817,892 bp	3,201,054 bp
assembly	contigs	12	108
assembly	N50	56,125 bp	1,568,056 bp
assembly	GC content	37%	38%
scaffolding	scaffolds	1	10
scaffolding	contigs	0	42
scaffolding	N50	657,549 bp	3,034,489 bp
annotation	coding genes	2,735	3,200
annotation	non-coding genes	95	144
antibiotic resistance	ABR genes	0	2
virulence factors	VF genes	16	35

https://doi.org/10.1371/journal.pcbi.1007134.t001 kmer-lookups as well as 16S ribosomal RNA database searches as *Listeria monocytogenes*, except of a single isolate classified as *Listeria innocua*. In line with these results all isolates shared an ANI value above 95% and a conserved DNA of at least 80% with at least one of the reference genomes, except for the *L. innocua* isolate which shared a maximum ANI of 90.7% and a conserved DNA of only 37.3%. Furthermore, the pipeline successfully subtyped all but one of the isolates via MLST, by automatically detecting and applying the "lmonocytogenes" schema. Noteworthy, the *L. innocua* isolate constitutes a distinct MLST lineage, *i.e. L. innocua*. ASA^{3}P detected between 0 and 2 antibiotic resistance genes and between 16 and 35 virulence factor genes. A comprehensive list of all key metrics for each genome is provided in a separate spreadsheet (S1 File).

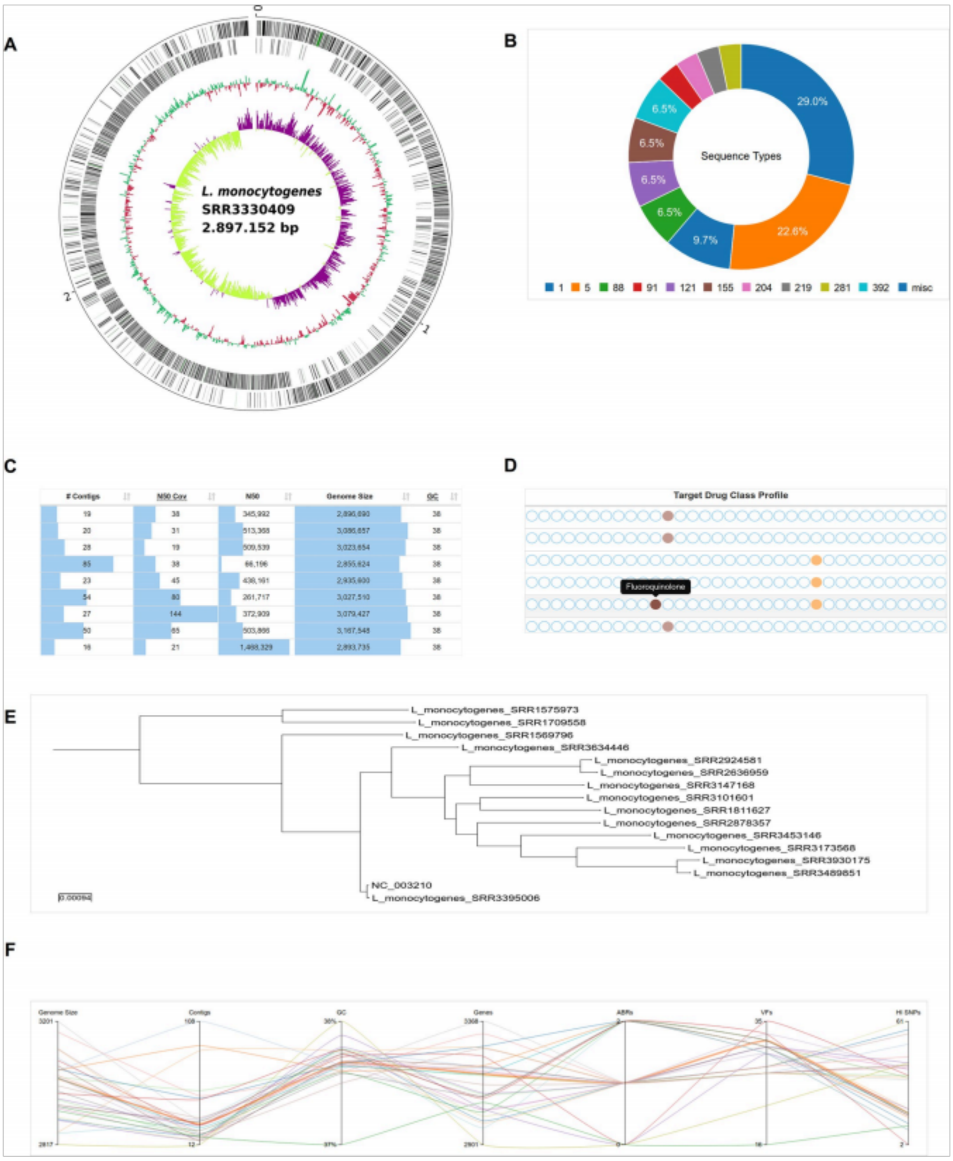
Finally, core and pan-genomes were computed resulting in 1,485 core genes and a pan-genome comprising 7,242 genes. Excluding the *L. innocua* strain and re-analyzing the dataset reduced the pan-genome to 6,197 genes and increased the amount of core genes to 2,004 additionally endorsing its taxonomic difference.

Data visualization

Analysis results as well as aggregated information get collected, transformed and finally presented by the pipeline via user friendly and detailed reports. These comprise local and responsive HTML5 documents containing interactive JavaScript visualizations facilitating the easy comprehension of the results. Fig 2 shows an exemplary collection of embedded data visualizations. Where appropriate, specialized widgets were implemented, as for instance circular genome annotation plots presenting genome features, GC content and GC skew on separate tracks (Fig 2A). These plots can be zoomed, panned and downloaded in SVG format for subsequent re-utilization. Another example is the interactive and dynamic visualization of SNP based phylogenetic trees (Fig 2E) via the Phylocanvas library (http://phylocanvas.org) enabling customizations by the user, as for instance changing tree types as well as collapsing and rotating subtrees. In order to provide users with an expeditious but conclusive overview on bacterial cohorts, key genome characteristics are visualized via an interactive parallel coordinates plot (Fig 2F) allowing for the combined selection of value ranges in different dimensions. Thus, clusters of isolates sharing high-level genome characteristics can be explored and identified straightforward. In order to rapidly compare different ABR capabilities of individual isolates, a specialized widget was designed and implemented (Fig 2D). For each isolate an ABR profile based on detected ABR genes grouped to 34 distinct target drug classes is computed, visualized and stacked for the easy perception of dissimilarities between genomes. Throughout the reports wherever appropriate, numeric results are interactively visualized as, for instance, the distribution of detected MLST sequence types (Fig 2B) and per-isolate analysis results summarized via key metrics presented within sortable and filterable data tables (Fig 2C).

Scalability and hardware requirements

When analyzing projects with growing numbers of isolates, local execution can quickly become infeasible. In order to address varying amounts of data, we provide two distinct ASA^{3}P distributions based on Docker and cloud computing environments. Each features individual scalability properties and implies different levels of technical complexity in terms of distribution and installation requirements. In order to benchmark the pipeline's scalability, we measured wall clock runtimes analyzing two projects comprising 32 and 1,024 *L. monocytogenes* isolates, respectively (S3 Table). Accession numbers for the large data set will be provided upon request. In addition to both public distributions, we also tested a custom installation on an inhouse SGE-based HPC cluster. The DV was executed on a VM providing



# Contigs	N50 Cov	N50	Genome Size	GC
19	38	345,992	2,896,690	38
20	31	513,368	3,086,657	38
28	19	509,539	3,023,654	38
85	38	66,196	2,855,624	38
23	45	436,161	2,935,600	38
54	80	261,717	3,027,510	38
27	144	372,909	3,079,427	38
50	65	503,866	3,167,548	38
16	21	1,488,329	2,893,735	38

Fig 2. Selection of interactive GUI widgets embedded in generated HTML5 reports. (A) Circular genome plot for a *Listeria monocytogenes* pseudogenome. The zoomable and scalable SVG based circular genome plot provides comprehensive information on genome features on mouseover events. Reference-guided rearranged contigs are linked to pseudogenomes for the sake of better readability. From the outermost inward: genes on the forward and reverse strand, respectively, GC content and GC skew. **(B)** Donut chart of MLST sequence type (ST) distribution. The MLST ST distribution of all isolates analyzed within a project is shown by and interactive donut chart. Single STs can be selected or deselected. **(C)** Visual representation of normalized assembly key statistics. Per-isolate assembly key statistics are normalized to minimum and maximum values within a project column-wise and visualized within an interactive data table allowing for column-based sorting and filtering for the rapid comparison of isolates and detection of outliers. **(D)** Antibiotic resistance profile overview widget. An antibiotic resistance profile comprising 34 distinct target drug classes is computed based on CARD annotations for each isolate and transformed into an overview widget allowing a rapid resistome comparison of all analyzed isolates. Black rectangle: a mouseover triggered tooltip describing detected antibiotic target drug resistance. **(E)** SNP-based approximately-maximum-likelihood phylogenetic tree. An approximately-maximum-likelihood phylogenetic tree is computed based on SNPs detected via read-mapping against a reference genome and stored in standard newick file format. The resulting tree is visualized via the interactive Phylocanvas JavaScript library providing comprehensive user interaction features, *e.g.* collapsing, expanding and rotating subtrees and tree type selection. **(F)** Parallel coordinates plot providing a multi-dimensional cohort overview of per-isolate genome metrics and characteristics. A selection of seven genome key metrics and characteristics is visualized in a parallel coordinates plot providing a multi-dimensional cohort overview enabling the rapid detection of clustered isolates and outliers. Vertical bars: key metrics or characteristic as plot dimensions; coloured horizontal lines: isolates and related values providing table-synchronized highlighting upon mouseovers.

https://doi.org/10.1371/journal.pcbi.1007134.g002

32 vCPUs and 64 GB memory. The quotas of the OS cloud project allowed for a total amount of 560 vCPUs and 1,280 GB memory. The HPC cluster comprised 20 machines with 40 cores and 256 GB memory, each. All machines hosted an Ubuntu 16.04 operating system. Table 2 shows the best-of-three runtimes for each version and benchmark data set combination. The pipeline successfully finished all benchmark analyses, except of the 1,024 dataset analyzed by the DV, due to lacking memory capacities required for the calculation of a phylogenetic tree comprising this large amount of genomes. Analyzing the 32 *L. monocytogenes* data set on larger compute infrastructures, *i.e.* the OS cloud (5:02:24 h) and HPC cluster (4:49:24 h), shows significantly reduced runtimes by approximately 50%, compared to the Docker-based executions (10:59:34 h). Not surprisingly, runtimes of the OSCV are slightly longer than HPC runtimes, due to the inherent overhead of automatic infrastructure setup and management procedures. Excluding these overheads reduces runtimes by approximately half an hour, leading to slightly shorter periods compared to the HPC version. We attribute this to a saturated workload distribution combined with faster CPUs in the cloud as stated in **S4 Table**. Comparing measured runtimes for both data sets exhibit a ~5.8- and ~6.9-fold increase for the HPC cluster (27:56:37 h) and OSCV (34:47:45 h) version, respectively, although the amount of isolates was increased 32-fold.

We furthermore investigated internal pipeline scaling properties for combinations of fixed and varying HPC cluster and project sizes (**S4 Fig**). In a first setup, growing numbers of *L. monocytogenes* isolates were analyzed utilizing a fixed-size HPC cluster of 4 compute nodes providing 32 vCPUs and 64 GB RAM each. Iteratively doubling the amount of isolates from 32 to 1,024 led to runtimes approximately increasing by a factor of 2, in line with our expectations. Nevertheless, we observed an overproportional increase in runtime of the internal comparative steps within stage C compared to the per-isolate steps of stage A and B. We attribute

Table 2. Wall clock runtimes for each ASA^{3}P version utilizing different hardware infrastructures and benchmark dataset sizes. Provided are best-of-three wall clock runtimes for complete ASA^{3}P executions analyzing *Listeria monocytogenes* benchmark datasets comprising 32 and 1,024 isolates given in hh:mm:ss format. Docker: a single virtual machine with 32 vCPUs and 64 GB memory was used. Analysis of the 1,024 isolate dataset was not feasible due to memory limitations; HPC: ASA^{3}P automatically distributed the workload to an SGE-based high-performance computing cluster comprising 20 nodes providing 40 cores and 256 GB memory each; Cloud: ASA^{3}P was executed in an OpenStack based cloud computing project comprising 560 vCPUs and 1,280 GB memory in total. Runtimes in parenthesis exclude build times for automatic infrastructure setups, *i.e.* the pure ASA^{3}P wall clock runtimes.

	Docker	Cloud	HPC
32 *L. monocytogenes*	10:59:34	5:02:24 (4:31:59)	4:49:24
1024 *L. monocytogenes*	-	34:47:45 (33:25:26)	27:56:37

https://doi.org/10.1371/journal.pcbi.1007134.t002

this to the implementations and inherent algorithms of internally used third party executables. As this might become a bottleneck for the analysis of even larger projects, this will be subject to future developments.

In addition, we repetitively analyzed a fixed number of 128 *L. monocytogenes* isolates while increasing underlying hardware capacities, *i.e.* available HPC compute nodes. In this second setup, we could measure significant runtime reductions for up to 8 compute nodes. Further hardware capacity expansions led to saturated workload distributions and contributed negligible runtime benefits. To summarize all conducted runtime benchmarks, we conclude that ASA^{3}P is able to horizontally scale-out to larger infrastructures and thus, conducting expeditious analysis of large projects within favourable periods of time.

To test the reliable distribution and robustness of the pipeline, we executed the DV on an Apple iMac running MacOS 10.14.2 providing 4 cores and 8 GB of memory. ASA^{3}P successfully analyzed a downsampled dataset comprising 4 *L. monocytogenes* isolates within a measured wall clock runtime of 8:43:12 hours. In order to assess minimal hardware requirements, the downsampled data set was analyzed iteratively reducing provided memory capacities of an OS VM. Hereby, we could determine a minimal memory requirement of 8 GB and thus draw the conclusion that ASA^{3}P allows the execution of a sophisticated workflow for the analysis of bacterial WGS data cohorts on ordinary consumer hardware. However, since larger amounts of isolates, more complex genomes or deeper sequencing coverages might result in higher hardware requirements, we nevertheless recommend at least 16 GB of memory.

Conclusion

We described ASA^{3}P, a new software tool for the local, automatic and highly scalable analysis of bacterial WGS data. The pipeline integrates many common analyses in a standardized and community best practices manner and is available for download either as a local command line tool encapsulated and distributed via Docker or a self-orchestrating OS cloud version. To the authors' best knowledge it is currently the only publicly available tool for the automatic high-throughput analysis of bacterial cohorts WGS data supporting all major contemporary sequencing platforms, offering SOPs, robust scalability as well as a user friendly and interactive graphical user interface whilst still being locally executable and thus offering on-premises analysis for sensitive or even confidential data. So far, ASA^{3}P has been used to analyze thousands of bacterial isolates covering a broad range of different taxa.

Availability and future directions

The source code is available on GitHub under GPL3 license at https://github.com/oschwengers/asap. The Docker container image is accessible at Docker Hub: https://hub.docker.com/r/oschwengers/asap. The ASA^{3}P software volume containing third-party executables and databases, OpenStack cloud scripts, a comprehensive manual and configuration templates are hosted at Zenodo: http://doi.org/10.5281/zenodo.3606300. Benchmark and exemplary data projects are hosted sepatately at Zenodo: https://doi.org/10.5281/zenodo.3606761. Questions and issues can be sent to "asap@computational.bio", bug reports can be filed as GitHub issues.

Albeit ASA^{3}P itself is published and distributed under a GPL3 license, some of its dependencies bundled within the ASA^{3}P volume are published under different license models, *e.g.* CARD and PubMLST. Comprehensive license information on each dependency and database is provided as a DEPENDENCY_LICENSE file within the ASA^{3}P directory.

Future directions comprise the development and integration of further analyses, *e.g.* detection and characterization of plasmids, phages and CRISPR cassettes as well as further enhancements in terms of scalability and usability.

Supporting information

S1 Table. Third party executable parameters and options. Parameters and options without scientific impact are excluded, *e.g.* input/output directories or number of threads.
(PDF)

S2 Table. Selection of task-specific third party bioinformatics software tools. Third party bioinformatics software tools selected for each task within ASA^{3}P along with a short argumentative reasoning for why they were selected.
(PDF)

S3 Table. Accession numbers of 32 *Listeria monocytogenes* isolates and reference genomes of the ASA^{3}P benchmark project. This exemplary project comprises 32 isolates from SRR Bioproject PRJNA215355 as well as two *Listeria monocytogenes* reference genomes from RefSeq. The project is provided as a GNU zipped tarball at http://doi.org/10.5281/zenodo.3606761
(PDF)

S4 Table. Host CPU information used for wall clock runtime benchmarks.
(PDF)

S1 Fig. Exemplary screenshot of configuration template sheet 1.
(PDF)

S2 Fig. Exemplary screenshot of configuration template sheet 2.
(PDF)

S3 Fig. Exemplary project directory structure. Each project analyzed by ASA^{3}P strictly follows a conventional directory organization and thus forestalls the burden of unnecessary configurations. Shown is an exemplary project structure representing input and output files and directories of the *Listeria monocytogenes* example project. For the sake of readability repeated blocks are collapsed represented by a triple dot '. . .'
(PDF)

S4 Fig. Wall clock runtimes for varying compute node and isolate numbers. Runtimes given in hours and separated between comparative and per-isolate internal pipeline stages due to different scalability metrics. Each compute node provides 32 vCPUs and 64 GB memory. *L. monocytogenes* strains were randomly chosen from SRA Bioproject PRJNA215355. **(A)** Runtimes of a fixed-size compute cluster comprising 4 compute nodes analyzing varying isolate numbers. **(B)** Runtimes of compute clusters with varying numbers of compute nodes analyzing a fixed amount of 128 isolates.
(PDF)

S1 File. Comprehensive list of all per-genome key metrics.
(XLS)

Acknowledgments

The authors thank the BiBiServ team from Bielefeld University for supply of the BiBiGrid framework and especially Jan Krüger for his comprehensive support. Furthermore, we thank Can Imirzalioglu, Jane Falgenhauer, Yancheng Yao and Swapnil Doijad at the Institute of Medical Microbiology of Justus-Liebig University Giessen for their ideas, bug reports and fruitful feedback on various aspects regarding implemented analyses and the graphical user interface.

Author Contributions

Conceptualization: Oliver Schwengers, Torsten Hain, Trinad Chakraborty, Alexander Goesmann.

Formal analysis: Oliver Schwengers.

Funding acquisition: Torsten Hain, Trinad Chakraborty, Alexander Goesmann.

Investigation: Oliver Schwengers.

Methodology: Oliver Schwengers, Moritz Fritzenwanker, Linda Falgenhauer, Torsten Hain.

Resources: Torsten Hain, Trinad Chakraborty, Alexander Goesmann.

Software: Oliver Schwengers, Andreas Hoek.

Validation: Moritz Fritzenwanker, Linda Falgenhauer, Torsten Hain.

Visualization: Oliver Schwengers, Andreas Hoek.

Writing – original draft: Oliver Schwengers.

Writing – review & editing: Linda Falgenhauer, Torsten Hain, Trinad Chakraborty, Alexander Goesmann.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977; 74: 5463–5467. https://doi.org/10.1073/pnas.74.12.5463 PMID: 271968
2. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014; 30: 418–426. https://doi.org/10.1016/j.tig.2014.07.001 PMID: 25108476
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016; 17: 333–351. https://doi.org/10.1038/nrg.2016.49 PMID: 27184599
4. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of *Mycoplasma genitalium*. Science. 1995; 270: 397–403. https://doi.org/10.1126/science.270.5235.397 PMID: 7569993
5. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995; 269: 496–512. https://doi.org/10.1126/science.7542800 PMID: 7542800
6. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 2018; 46: D851–D860. https://doi.org/10.1093/nar/gkx1068 PMID: 29112715
7. Long SW, Williams D, Valson C, Cantu CC, Cernoch P, Musser JM, et al. A genomic day in the life of a clinical microbiology laboratory. J Clin Microbiol. 2013; 51: 1272–1277. https://doi.org/10.1128/JCM.03237-12 PMID: 23345298
8. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012; 13: 601–612. https://doi.org/10.1038/nrg3226 PMID: 22868263
9. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." Proceedings of the National Academy of Sciences. 2005; 102: 13950–13955.
10. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol. Elsevier; 2017; 243: 16–24. https://doi.org/10.1016/j.jbiotec.2016.12.022 PMID: 28042011
11. Review on Antimicrobial Resistance. Tackling Drug-Resistant Infections Globally: final report and recommendations [Internet]. Wellcome Trust; 2016 May. Available: https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf
12. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, ECDC National Microbiology Focal Points and Experts Group ENMFPAE. Survey on the Use of Whole-Genome Sequencing for Infectious

Diseases Surveillance: Rapid Expansion of European National Capacities, 2015–2016. Frontiers in public health. Frontiers Media SA; 2017; 5: 347. https://doi.org/10.3389/fpubh.2017.00347 PMID: 29326921

13. Glaser P, Martins-Simões P, Villain A, Barbier M, Tristan A, Bouchier C, et al. Demography and Intercontinental Spread of the USA300 Community-Acquired Methicillin-Resistant *Staphylococcus aureus* Lineage. MBio. 2016; 7: e02183–15. https://doi.org/10.1128/mBio.02183-15 PMID: 26884428

14. Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. Genome Res. 2013; 23: 653–664. https://doi.org/10.1101/gr.147710.112 PMID: 23299977

15. Nübel U. Emergence and Spread of Antimicrobial Resistance: Recent Insights from Bacterial Population Genomics. Curr Top Microbiol Immunol. 398: 35–53. https://doi.org/10.1007/82_2016_505 PMID: 27738914

16. Baur D, Gladstone BP, Burkert F, Carrara E, Foschi F, Döbele S, et al. Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and *Clostridium difficile* infection: a systematic review and meta-analysis. Lancet Infect Dis. 2017; 17: 990–1001. https://doi.org/10.1016/S1473-3099(17)30325-0 PMID: 28629876

17. Schempp FM, Drummond L, Buchhaupt M, Schrader J. Microbial Cell Factories for the Production of Terpenoid Flavor and Fragrance Compounds. J Agric Food Chem. 2018; 66: 2247–2258. https://doi.org/10.1021/acs.jafc.7b00473 PMID: 28418659

18. Corchero JL, Gasser B, Resina D, Smith W, Parrilli E, Vázquez F, et al. Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics. Biotechnol Adv. 31: 140–153. https://doi.org/10.1016/j.biotechadv.2012.09.001 PMID: 22985698

19. Huang C-J, Lin H, Yang X. Industrial production of recombinant therapeutics in *Escherichia coli* and its recent advancements. J Ind Microbiol Biotechnol. 2012; 39: 383–399. https://doi.org/10.1007/s10295-011-1082-9 PMID: 22252444

20. Baeshen MN, Al-Hejin AM, Bora RS, Ahmed MMM, Ramadan HAI, Saini KS, et al. Production of Biopharmaceuticals in E. coli: Current Scenario and Future Perspectives. J Microbiol Biotechnol. 2015; 25: 953–962. https://doi.org/10.4014/jmb.1412.12079 PMID: 25737124

21. Wackett LP. Microbial-based motor fuels: science and technology. Microb Biotechnol. 2008; 1: 211–225. https://doi.org/10.1111/j.1751-7915.2007.00020.x PMID: 21261841

22. Zhang W, Yin K, Chen L. Bacteria-mediated bisphenol A degradation. Appl Microbiol Biotechnol. 2013; 97: 5681–5689. https://doi.org/10.1007/s00253-013-4949-z PMID: 23681588

23. Singh B, Kaur J, Singh K. Microbial remediation of explosive waste. Crit Rev Microbiol. 2012; 38: 152–167. https://doi.org/10.3109/1040841X.2011.640979 PMID: 22497284

24. Kip N, van Veen JA. The dual role of microbes in corrosion. ISME J. 2015; 9: 542–551. https://doi.org/10.1038/ismej.2014.169 PMID: 25259571

25. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015; 13: e1002195. https://doi.org/10.1371/journal.pbio.1002195 PMID: 26151137

26. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol. Genome Biology; 2016; 17: 1–9. https://doi.org/10.1186/s13059-015-0866-z

27. Gargis AS, Kalman L, Lubin IM. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol. 2016; 54: 2857–2865. https://doi.org/10.1128/JCM.00949-16 PMID: 27510831

28. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008; 9: 75. https://doi.org/10.1186/1471-2164-9-75 PMID: 18261238

29. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017; 45: D535–D542. https://doi.org/10.1093/nar/gkw1017 PMID: 27899627

30. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014; 30: 2068–2069. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

31. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

32. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

33. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. Nature Publishing Group, a

division of Macmillan Publishers Limited. All Rights Reserved.; 2013; 10: 563–569. https://doi.org/10.1038/nmeth.2474 PMID: 23644548

34. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLoS Comput Biol. Public Library of Science; 2017; 13: e1005595. https://doi.org/10.1371/journal.pcbi.1005595 PMID: 28594827

35. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, et al. MeDuSa: A multi-draft based scaffolder. Bioinformatics. 2015; 31: 2443–2451. https://doi.org/10.1093/bioinformatics/btv171 PMID: 25810435

36. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017; 45: D566–D573. https://doi.org/10.1093/nar/gkw1004 PMID: 27789705

37. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: Hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res. 2016; 44: D694–D697. https://doi.org/10.1093/nar/gkv1239 PMID: 26578559

38. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007; 57: 81–91. https://doi.org/10.1099/ijs.0.64483-0 PMID: 17220447

39. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15: R46. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807

40. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

41. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013; 41: D590–6. https://doi.org/10.1093/nar/gks1219 PMID: 23193283

42. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5: R12. https://doi.org/10.1186/gb-2004-5-2-r12 PMID: 14759262

43. Jolley KA, Bray JE, Maiden MCJ. A RESTful application programming interface for the PubMLST molecular typing and genome databases. Database. 2017;2017. https://doi.org/10.1093/database/bax060 PMID: 29220452

44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

45. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018; 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191 PMID: 29750242

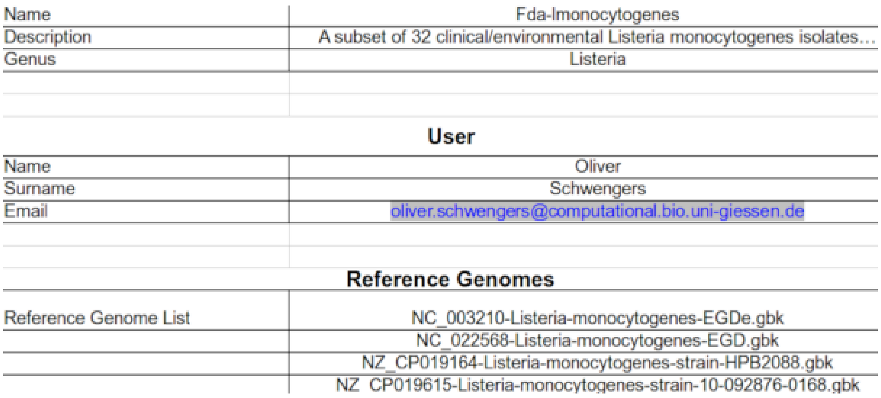
46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

47. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012; 6: 80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672

48. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One. 2010; 5: e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

49. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015; 31: 3691–3693. https://doi.org/10.1093/bioinformatics/btv421 PMID: 26198102

7.2 Supplementary Information – ASA³P



Name	Fda-lmonocytogenes
Description	A subset of 32 clinical/environmental Listeria monocytogenes isolates…
Genus	Listeria

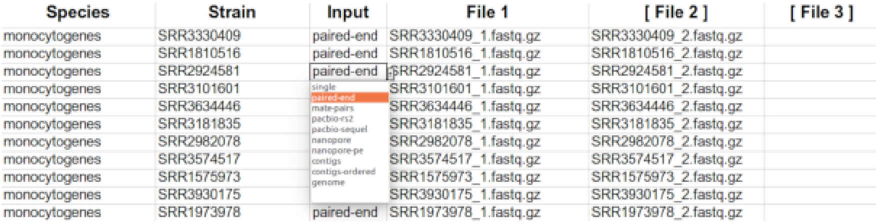
User

Name	Oliver
Surname	Schwengers
Email	oliver.schwengers@computational.bio.uni-giessen.de

Reference Genomes

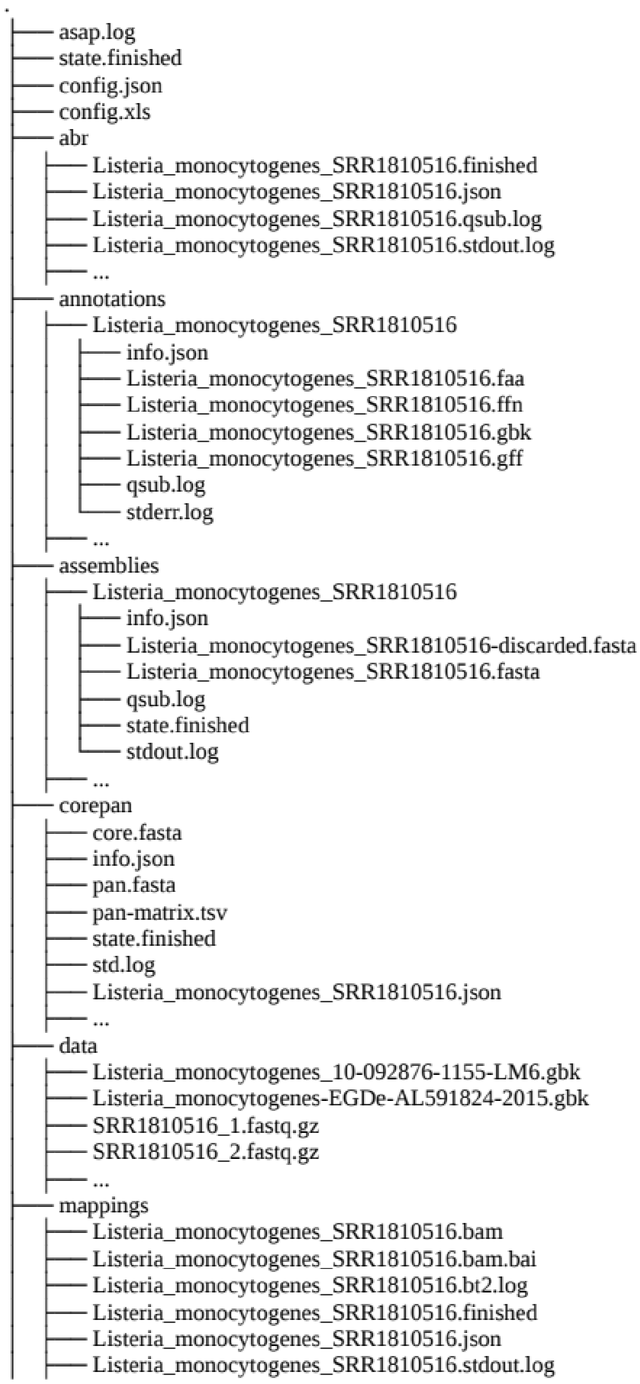
Reference Genome List	NC_003210-Listeria-monocytogenes-EGDe.gbk
	NC_022568-Listeria-monocytogenes-EGD.gbk
	NZ_CP019164-Listeria-monocytogenes-strain-HPB2088.gbk
	NZ CP019615-Listeria-monocytogenes-strain-10-092876-0168.gbk

S1 Fig. Exemplary screenshot of configuration template sheet 1.

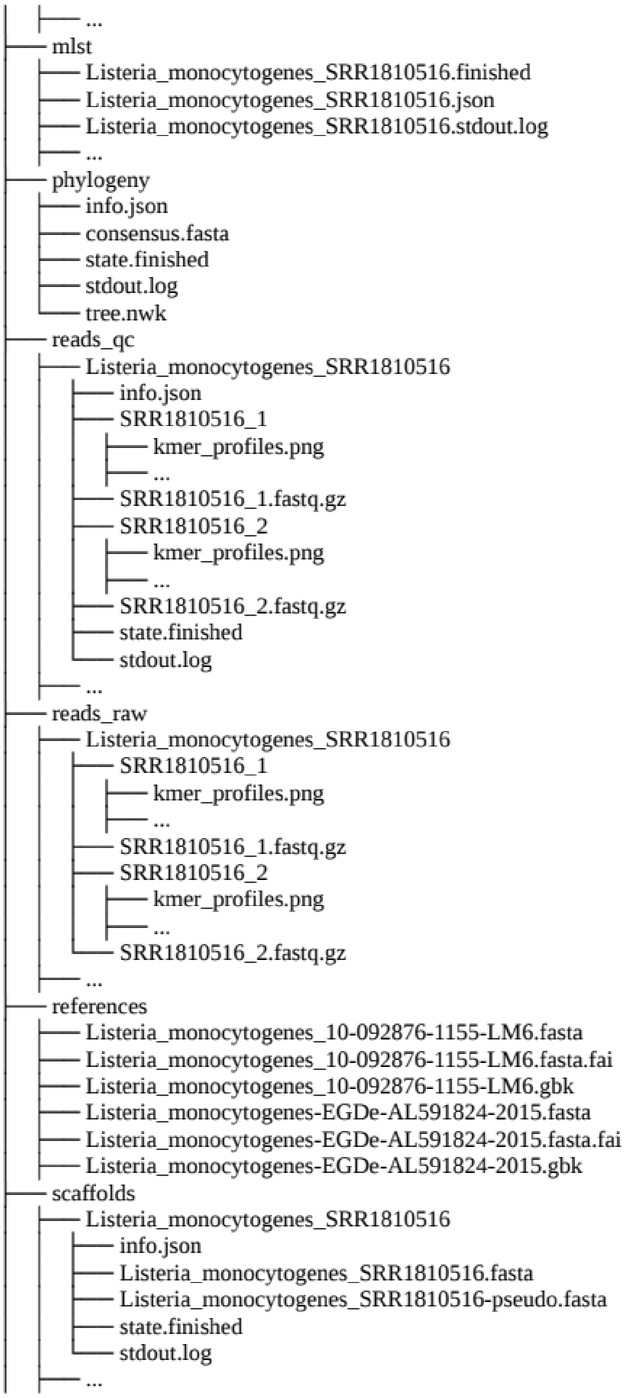


Species	Strain	Input	File 1	[File 2]	[File 3]
monocytogenes	SRR3330409	paired-end	SRR3330409_1.fastq.gz	SRR3330409_2.fastq.gz	
monocytogenes	SRR1810516	paired-end	SRR1810516_1.fastq.gz	SRR1810516_2.fastq.gz	
monocytogenes	SRR2924581	paired-end	SRR2924581_1.fastq.gz	SRR2924581_2.fastq.gz	
monocytogenes	SRR3101601		SRR3101601_1.fastq.gz	SRR3101601_2.fastq.gz	
monocytogenes	SRR3634446		SRR3634446_1.fastq.gz	SRR3634446_2.fastq.gz	
monocytogenes	SRR3181835		SRR3181835_1.fastq.gz	SRR3181835_2.fastq.gz	
monocytogenes	SRR2982078		SRR2982078_1.fastq.gz	SRR2982078_2.fastq.gz	
monocytogenes	SRR3574517		SRR3574517_1.fastq.gz	SRR3574517_2.fastq.gz	
monocytogenes	SRR1575973		SRR1575973_1.fastq.gz	SRR1575973_2.fastq.gz	
monocytogenes	SRR3930175		SRR3930175_1.fastq.gz	SRR3930175_2.fastq.gz	
monocytogenes	SRR1973978	paired-end	SRR1973978_1.fastq.gz	SRR1973978_2.fastq.gz	

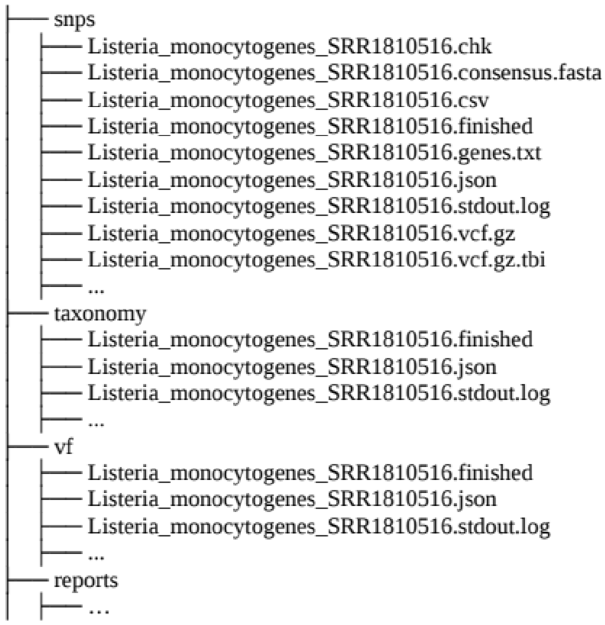
S2 Fig. Exemplary screenshot of configuration template sheet 2.



```
.
├── asap.log
├── state.finished
├── config.json
├── config.xls
├── abr
│   ├── Listeria_monocytogenes_SRR1810516.finished
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── Listeria_monocytogenes_SRR1810516.qsub.log
│   ├── Listeria_monocytogenes_SRR1810516.stdout.log
│   ├── ...
├── annotations
│   ├── Listeria_monocytogenes_SRR1810516
│   │   ├── info.json
│   │   ├── Listeria_monocytogenes_SRR1810516.faa
│   │   ├── Listeria_monocytogenes_SRR1810516.ffn
│   │   ├── Listeria_monocytogenes_SRR1810516.gbk
│   │   ├── Listeria_monocytogenes_SRR1810516.gff
│   │   ├── qsub.log
│   │   └── stderr.log
│   ├── ...
├── assemblies
│   ├── Listeria_monocytogenes_SRR1810516
│   │   ├── info.json
│   │   ├── Listeria_monocytogenes_SRR1810516-discarded.fasta
│   │   ├── Listeria_monocytogenes_SRR1810516.fasta
│   │   ├── qsub.log
│   │   ├── state.finished
│   │   └── stdout.log
│   ├── ...
├── corepan
│   ├── core.fasta
│   ├── info.json
│   ├── pan.fasta
│   ├── pan-matrix.tsv
│   ├── state.finished
│   ├── std.log
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── ...
├── data
│   ├── Listeria_monocytogenes_10-092876-1155-LM6.gbk
│   ├── Listeria_monocytogenes-EGDe-AL591824-2015.gbk
│   ├── SRR1810516_1.fastq.gz
│   ├── SRR1810516_2.fastq.gz
│   ├── ...
├── mappings
│   ├── Listeria_monocytogenes_SRR1810516.bam
│   ├── Listeria_monocytogenes_SRR1810516.bam.bai
│   ├── Listeria_monocytogenes_SRR1810516.bt2.log
│   ├── Listeria_monocytogenes_SRR1810516.finished
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── Listeria_monocytogenes_SRR1810516.stdout.log
```

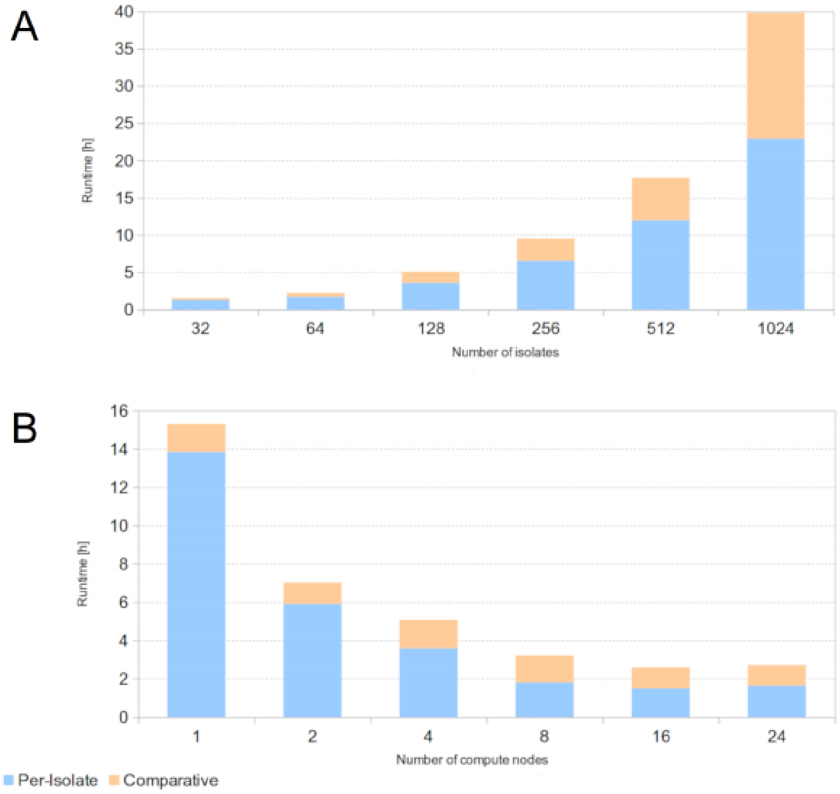


```
│   ├── ...
├── mlst
│   ├── Listeria_monocytogenes_SRR1810516.finished
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── Listeria_monocytogenes_SRR1810516.stdout.log
│   ├── ...
├── phylogeny
│   ├── info.json
│   ├── consensus.fasta
│   ├── state.finished
│   ├── stdout.log
│   └── tree.nwk
├── reads_qc
│   ├── Listeria_monocytogenes_SRR1810516
│   │   ├── info.json
│   │   ├── SRR1810516_1
│   │   │   ├── kmer_profiles.png
│   │   │   ├── ...
│   │   ├── SRR1810516_1.fastq.gz
│   │   ├── SRR1810516_2
│   │   │   ├── kmer_profiles.png
│   │   │   ├── ...
│   │   ├── SRR1810516_2.fastq.gz
│   │   ├── state.finished
│   │   └── stdout.log
│   ├── ...
├── reads_raw
│   ├── Listeria_monocytogenes_SRR1810516
│   │   ├── SRR1810516_1
│   │   │   ├── kmer_profiles.png
│   │   │   ├── ...
│   │   ├── SRR1810516_1.fastq.gz
│   │   ├── SRR1810516_2
│   │   │   ├── kmer_profiles.png
│   │   │   ├── ...
│   │   └── SRR1810516_2.fastq.gz
│   ├── ...
├── references
│   ├── Listeria_monocytogenes_10-092876-1155-LM6.fasta
│   ├── Listeria_monocytogenes_10-092876-1155-LM6.fasta.fai
│   ├── Listeria_monocytogenes_10-092876-1155-LM6.gbk
│   ├── Listeria_monocytogenes-EGDe-AL591824-2015.fasta
│   ├── Listeria_monocytogenes-EGDe-AL591824-2015.fasta.fai
│   ├── Listeria_monocytogenes-EGDe-AL591824-2015.gbk
├── scaffolds
│   ├── Listeria_monocytogenes_SRR1810516
│   │   ├── info.json
│   │   ├── Listeria_monocytogenes_SRR1810516.fasta
│   │   ├── Listeria_monocytogenes_SRR1810516-pseudo.fasta
│   │   ├── state.finished
│   │   └── stdout.log
│   ├── ...
```

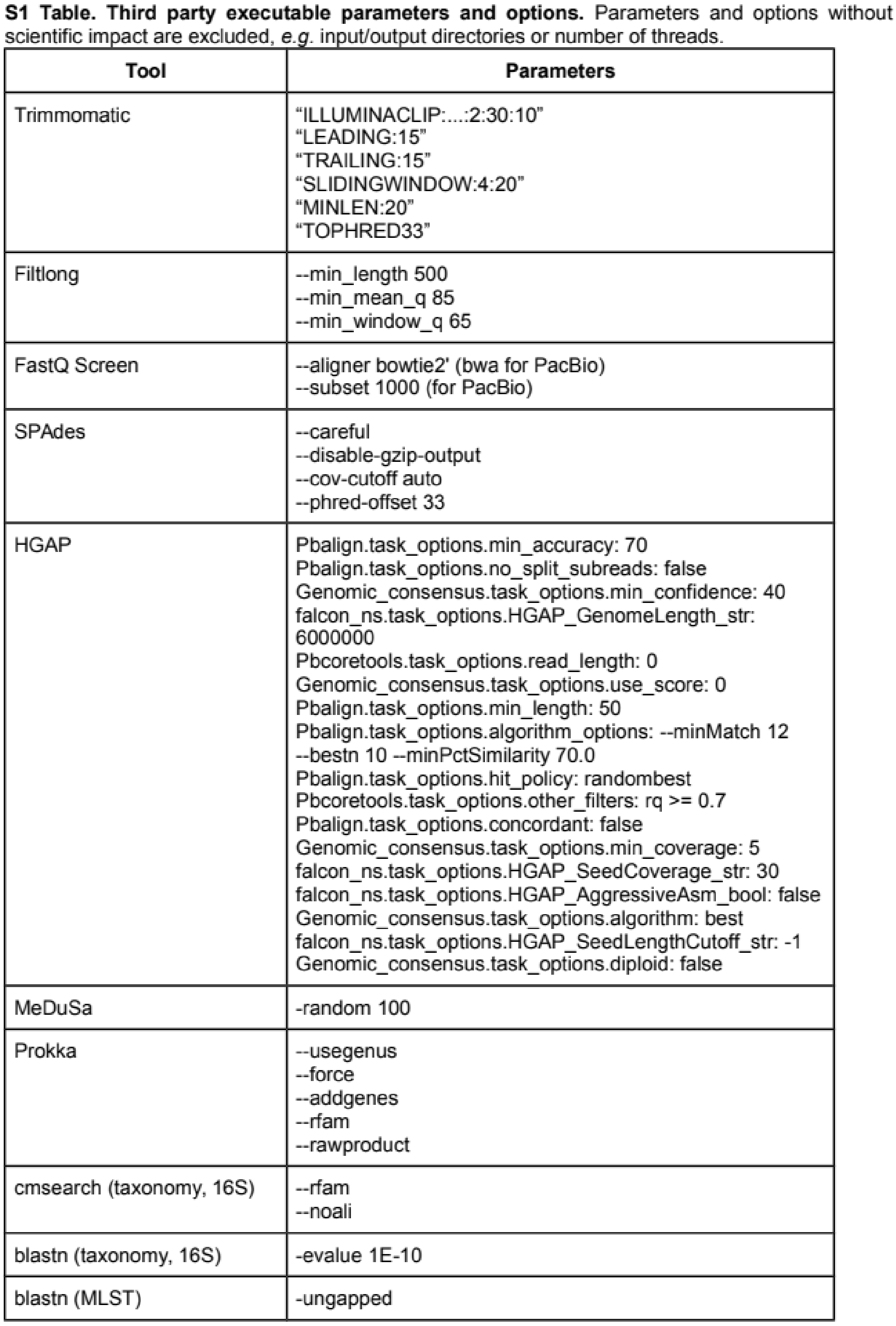


```
├── snps
│   ├── Listeria_monocytogenes_SRR1810516.chk
│   ├── Listeria_monocytogenes_SRR1810516.consensus.fasta
│   ├── Listeria_monocytogenes_SRR1810516.csv
│   ├── Listeria_monocytogenes_SRR1810516.finished
│   ├── Listeria_monocytogenes_SRR1810516.genes.txt
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── Listeria_monocytogenes_SRR1810516.stdout.log
│   ├── Listeria_monocytogenes_SRR1810516.vcf.gz
│   ├── Listeria_monocytogenes_SRR1810516.vcf.gz.tbi
│   ├── ...
├── taxonomy
│   ├── Listeria_monocytogenes_SRR1810516.finished
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── Listeria_monocytogenes_SRR1810516.stdout.log
│   ├── ...
├── vf
│   ├── Listeria_monocytogenes_SRR1810516.finished
│   ├── Listeria_monocytogenes_SRR1810516.json
│   ├── Listeria_monocytogenes_SRR1810516.stdout.log
│   ├── ...
├── reports
│   ├── …
```

S3 Fig. Exemplary project directory structure. Each project analyzed by ASA³P strictly follows a conventional directory organization and thus forestalls the burden of unnecessary configurations. Shown is an exemplary project structure representing input and output files and directories of the *Listeria monocytogenes* example project. For the sake of readability repeated blocks are collapsed represented by a triple dot '...'

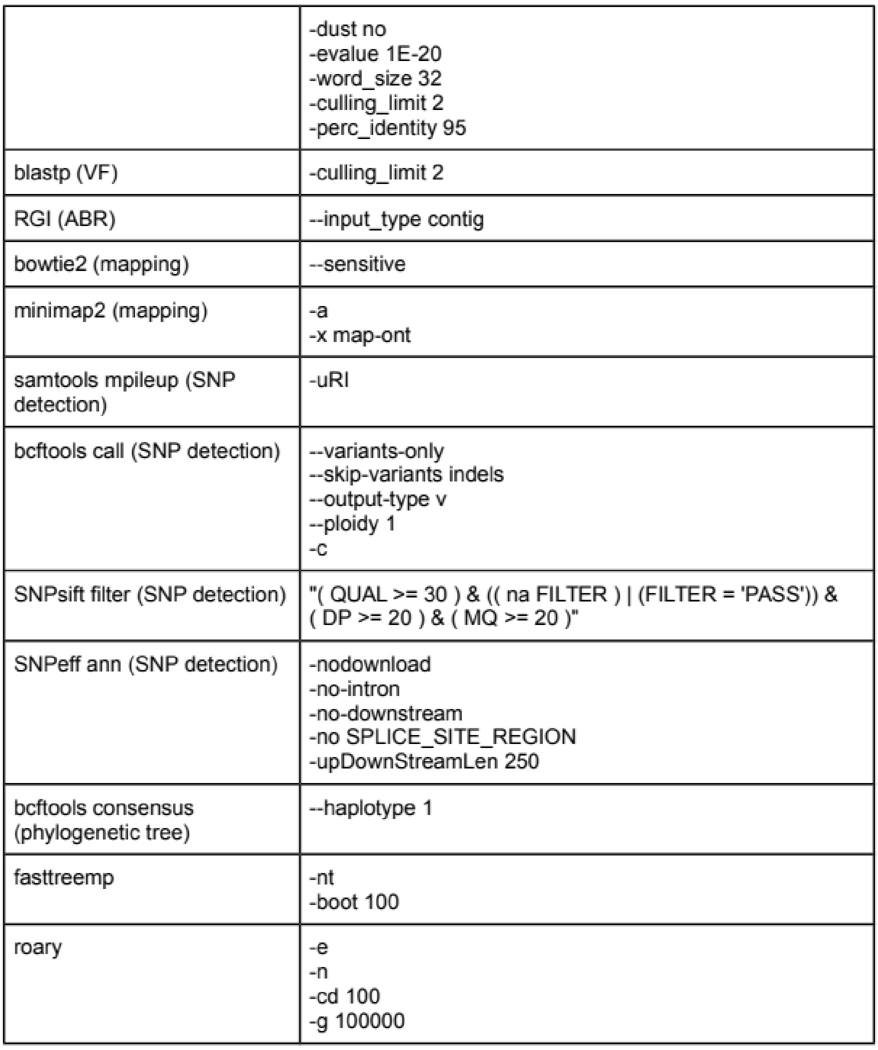


S4 Fig. Wall clock runtimes for varying compute node and isolate numbers. Runtimes given in hours and separated between comparative and per-isolate internal pipeline stages due to different scalability metrics. Each compute node provides 32 vCPUs and 64 GB memory. *L. monocytogenes* strains were randomly chosen from SRA Bioproject PRJNA215355. (A) Runtimes of a fixed-size compute cluster comprising 4 compute nodes analyzing varying isolate numbers. (B**)** Runtimes of compute clusters with varying numbers of compute nodes analyzing a fixed amount of 128 isolates.

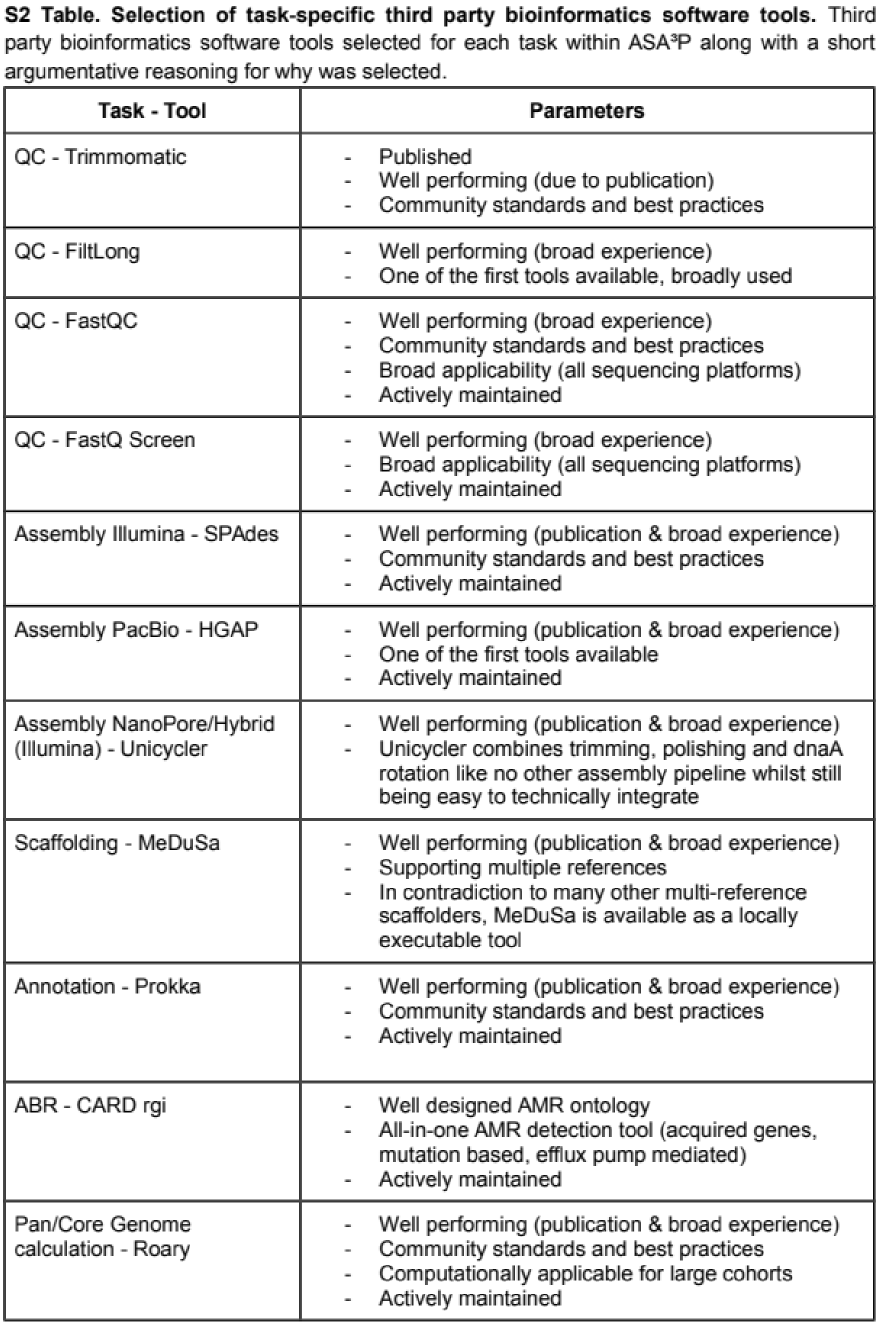


S1 Table. Third party executable parameters and options. Parameters and options without scientific impact are excluded, *e.g.* input/output directories or number of threads.

Tool	Parameters
Trimmomatic	“ILLUMINACLIP:...:2:30:10” “LEADING:15” “TRAILING:15” “SLIDINGWINDOW:4:20” “MINLEN:20” “TOPHRED33”
Filtlong	--min_length 500 --min_mean_q 85 --min_window_q 65
FastQ Screen	--aligner bowtie2' (bwa for PacBio) --subset 1000 (for PacBio)
SPAdes	--careful --disable-gzip-output --cov-cutoff auto --phred-offset 33
HGAP	Pbalign.task_options.min_accuracy: 70 Pbalign.task_options.no_split_subreads: false Genomic_consensus.task_options.min_confidence: 40 falcon_ns.task_options.HGAP_GenomeLength_str: 6000000 Pbcoretools.task_options.read_length: 0 Genomic_consensus.task_options.use_score: 0 Pbalign.task_options.min_length: 50 Pbalign.task_options.algorithm_options: --minMatch 12 --bestn 10 --minPctSimilarity 70.0 Pbalign.task_options.hit_policy: randombest Pbcoretools.task_options.other_filters: rq >= 0.7 Pbalign.task_options.concordant: false Genomic_consensus.task_options.min_coverage: 5 falcon_ns.task_options.HGAP_SeedCoverage_str: 30 falcon_ns.task_options.HGAP_AggressiveAsm_bool: false Genomic_consensus.task_options.algorithm: best falcon_ns.task_options.HGAP_SeedLengthCutoff_str: -1 Genomic_consensus.task_options.diploid: false
MeDuSa	-random 100
Prokka	--usegenus --force --addgenes --rfam --rawproduct
cmsearch (taxonomy, 16S)	--rfam --noali
blastn (taxonomy, 16S)	-evalue 1E-10
blastn (MLST)	-ungapped

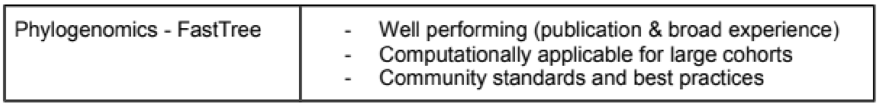


	-dust no -evalue 1E-20 -word_size 32 -culling_limit 2 -perc_identity 95
blastp (VF)	-culling_limit 2
RGI (ABR)	--input_type contig
bowtie2 (mapping)	--sensitive
minimap2 (mapping)	-a -x map-ont
samtools mpileup (SNP detection)	-uRI
bcftools call (SNP detection)	--variants-only --skip-variants indels --output-type v --ploidy 1 -c
SNPsift filter (SNP detection)	"(QUAL >= 30) & ((na FILTER) | (FILTER = 'PASS')) & (DP >= 20) & (MQ >= 20)"
SNPeff ann (SNP detection)	-nodownload -no-intron -no-downstream -no SPLICE_SITE_REGION -upDownStreamLen 250
bcftools consensus (phylogenetic tree)	--haplotype 1
fasttreemp	-nt -boot 100
roary	-e -n -cd 100 -g 100000



S2 Table. Selection of task-specific third party bioinformatics software tools. Third party bioinformatics software tools selected for each task within ASA^{3}P along with a short argumentative reasoning for why was selected.

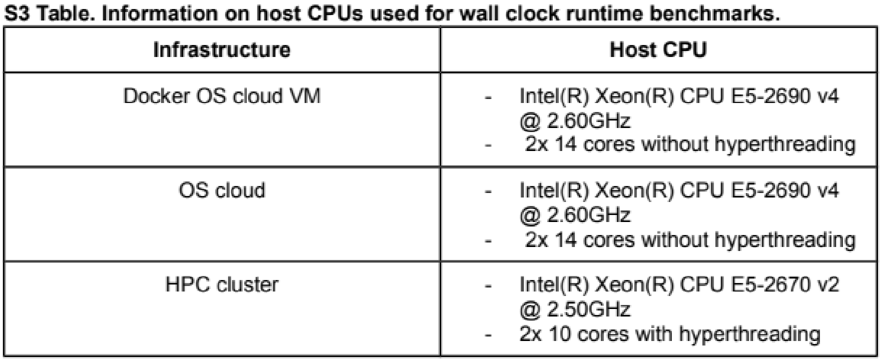
Task - Tool	Parameters
QC - Trimmomatic	- Published - Well performing (due to publication) - Community standards and best practices
QC - FiltLong	- Well performing (broad experience) - One of the first tools available, broadly used
QC - FastQC	- Well performing (broad experience) - Community standards and best practices - Broad applicability (all sequencing platforms) - Actively maintained
QC - FastQ Screen	- Well performing (broad experience) - Broad applicability (all sequencing platforms) - Actively maintained
Assembly Illumina - SPAdes	- Well performing (publication & broad experience) - Community standards and best practices - Actively maintained
Assembly PacBio - HGAP	- Well performing (publication & broad experience) - One of the first tools available - Actively maintained
Assembly NanoPore/Hybrid (Illumina) - Unicycler	- Well performing (publication & broad experience) - Unicycler combines trimming, polishing and dnaA rotation like no other assembly pipeline whilst still being easy to technically integrate
Scaffolding - MeDuSa	- Well performing (publication & broad experience) - Supporting multiple references - In contradiction to many other multi-reference scaffolders, MeDuSa is available as a locally executable tool
Annotation - Prokka	- Well performing (publication & broad experience) - Community standards and best practices - Actively maintained
ABR - CARD rgi	- Well designed AMR ontology - All-in-one AMR detection tool (acquired genes, mutation based, efflux pump mediated) - Actively maintained
Pan/Core Genome calculation - Roary	- Well performing (publication & broad experience) - Community standards and best practices - Computationally applicable for large cohorts - Actively maintained



Phylogenomics - FastTree	- Well performing (publication & broad experience) - Computationally applicable for large cohorts - Community standards and best practices

S2 Table. Accession numbers of 32 *Listeria monocytogenes* isolates and reference genomes of the ASA³P benchmark project. This exemplary project comprises 32 isolates from SRR Bioproject PRJNA215355 as well as two *Listeria monocytogenes* reference genomes from RefSeq. The project is provided as a GNU zipped tarball at https://s3.computational.bio.uni-giesen.de/swift/v1/asap/example-lmonocytogenes-32.tar.gz

Type	Accession numbers
reference	NC_003210.1 NC_022568.1 NZ_CP019164.1 NZ_CP019615.1
isolates	SRR3330409, SRR1810516, SRR2924581, SRR3101601, SRR3634446, SRR3181835, SRR2982078, SRR3574517, SRR1575973, SRR3930175, SRR1973978, SRR2140707, SRR2976738, SRR2636959, SRR3173568, SRR3928673, SRR1709558, SRR1514752, SRR3489851, SRR1811627, SRR2878357, SRR1272887, SRR3147168, SRR2533768, SRR1569796, SRR1763858, SRR3395006, SRR3930198, SRR2861532, SRR2562281, SRR3453146, SRR3137565



S3 Table. Information on host CPUs used for wall clock runtime benchmarks.

Infrastructure	Host CPU
Docker OS cloud VM	- Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz - 2x 14 cores without hyperthreading
OS cloud	- Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz - 2x 14 cores without hyperthreading
HPC cluster	- Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz - 2x 10 cores with hyperthreading

7.3 Platon: identification and characterization of bacterial plasmid contigs from short-read draft assemblies exploiting protein-sequence-based replicon distribution scores.

Oliver Schwengers[1,2,3 *], Patrick Barth[1], Linda Falgenhauer[2,3], Torsten Hain[2,3], Trinad Chakraborty[2,3,‡], Alexander Goesmann[1,3,‡]

Affiliations

[1] Bioinformatics and Systems Biology,
Justus Liebig University Giessen, Giessen, Germany

[2] Institute of Medical Microbiology,
Justus Liebig University Giessen, Giessen, Germany

[3] German Center for Infection Research (DZIF),
partner site Giessen-Marburg-Langen, Giessen, Germany

* Corresponding author
‡ These authors contributed equally to this work.

Journal Microbial Genomics

Date 2020

Volume 6(10)

DOI 10.1099/mgen.0.000398

7.3 Platon: identification and characterization of bacterial plasmid contigs from short-read draft assemblies exploiting protein-sequence-based replicon distribution scores.

MICROBIAL GENOMICS

RESEARCH ARTICLE
Schwengers *et al., Microbial Genomics*
DOI 10.1099/mgen.0.000398

Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores

Oliver Schwengers[1,2,3,*], Patrick Barth[1], Linda Falgenhauer[2,3,‡], Torsten Hain[2,3], Trinad Chakraborty[2,3,†] and Alexander Goesmann[1,3,†]

Abstract

Plasmids are extrachromosomal genetic elements that replicate independently of the chromosome and play a vital role in the environmental adaptation of bacteria. Due to potential mobilization or conjugation capabilities, plasmids are important genetic vehicles for antimicrobial resistance genes and virulence factors with huge and increasing clinical implications. They are therefore subject to large genomic studies within the scientific community worldwide. As a result of rapidly improving next-generation sequencing methods, the quantity of sequenced bacterial genomes is constantly increasing, in turn raising the need for specialized tools to (i) extract plasmid sequences from draft assemblies, (ii) derive their origin and distribution, and (iii) further investigate their genetic repertoire. Recently, several bioinformatic methods and tools have emerged to tackle this issue; however, a combination of high sensitivity and specificity in plasmid sequence identification is rarely achieved in a taxon-independent manner. In addition, many software tools are not appropriate for large high-throughput analyses or cannot be included in existing software pipelines due to their technical design or software implementation. In this study, we investigated differences in the replicon distributions of protein-coding genes on a large scale as a new approach to distinguish plasmid-borne from chromosome-borne contigs. We defined and computed statistical discrimination thresholds for a new metric: the replicon distribution score (RDS), which achieved an accuracy of 96.6%. The final performance was further improved by the combination of the RDS metric with heuristics exploiting several plasmid-specific higher-level contig characterizations. We implemented this workflow in a new high-throughput taxon-independent bioinformatics software tool called Platon for the recruitment and characterization of plasmid-borne contigs from short-read draft assemblies. Compared to PlasFlow, Platon achieved a higher accuracy (97.5%) and more balanced predictions (F1=82.6%) tested on a broad range of bacterial taxa and better or equal performance against the targeted tools PlasmidFinder and PlaScope on sequenced *Escherichia coli* isolates. Platon is available at: http://platon.computational.bio/.

DATA SUMMARY

(1) Platon was developed as a Python 3 command line application for Linux.

(2) The complete source code and documentation are available on GitHub under a GPL3 license: https://github.com/oschwengers/platon and http://platon.computational.bio.

(3) All database versions are hosted at Zenodo (DOI: 10.5281/zenodo.3349651).

(4) Platon is available via the bioconda package platon.

(5) Platon is available via the PyPI package cb-platon.

(6) The bacterial representative sequences for UniProt's UniRef90 protein clusters, complete bacterial genome

Received 03 December 2019; Accepted 02 June 2020; Published 24 June 2020

Author affiliations: [1]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany; [2]Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, Germany; [3]German Center for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Giessen, Germany.
***Correspondence:** Oliver Schwengers, oliver.schwengers@computational.bio.uni-giessen.de

Keywords: bacteria; plasmids; NGS; whole-genome sequencing.

Abbreviations: CDS, coding sequence; CT, conservative threshold; HMM, hidden Markov model; MCC, Matthews correlation coefficient; MPS, marker protein sequence; NCBI, National Center for Biotechnology Information; NPV, negative predictive value; PPV, positive predictive value; RDS, replicon distribution score; SNT, sensitivity threshold; SP, specificity threshold; T4CP, type4-coupling protein.

‡Present address: Institute of Hygiene and Environmental Health, Justus Liebig University, Giessen, Germany.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary tables and two supplementary figures are available with the online version of this article.

sequences from the National Center for Biotechnology Information (NCBI) RefSeq database, complete plasmid sequences from the NCBI genomes plasmid section, created artificial contigs, replicon distribution score (RDS) threshold metrics and raw protein replicon hit counts used to create and evaluate the marker protein sequence database are hosted at Zenodo (DOI: 10.5281/zenodo.3759169).

(7) Twenty-four *Escherichia coli* isolates sequenced with short-read (Illumina MiSeq) and long-read sequencing technologies (Oxford Nanopore Technology GridION platform) used for real data benchmarks are available under the following NCBI BioProjects: PRJNA505407 and PRJNA387731.

INTRODUCTION

Plasmids are bacterial extrachromosomal DNA elements that replicate independently of the chromosome. They are mostly circular, have characteristic copy numbers per cell and carry genes that are usually not essential under normal conditions but rather allow bacteria to adapt to specific environments and conditions [1]. These genes, for instance, provide antibiotic or heavy metal resistance, are involved in alternative metabolic pathways or encode for virulence factors [2]. As plasmids are not only inherited by daughter cells, but can also be dispersed by horizontal gene transfer, they can spread rapidly within and between bacterial populations [3–5]. For example, identical antibiotic resistance plasmids have already been isolated from humans and animals [6]. Thus, plasmids are important mediators of antibiotic resistance spread and recent findings have confirmed that they frequently play a major role in clinical outbreaks [7, 8]. Therefore, it is of huge importance to properly identify and analyse plasmids.

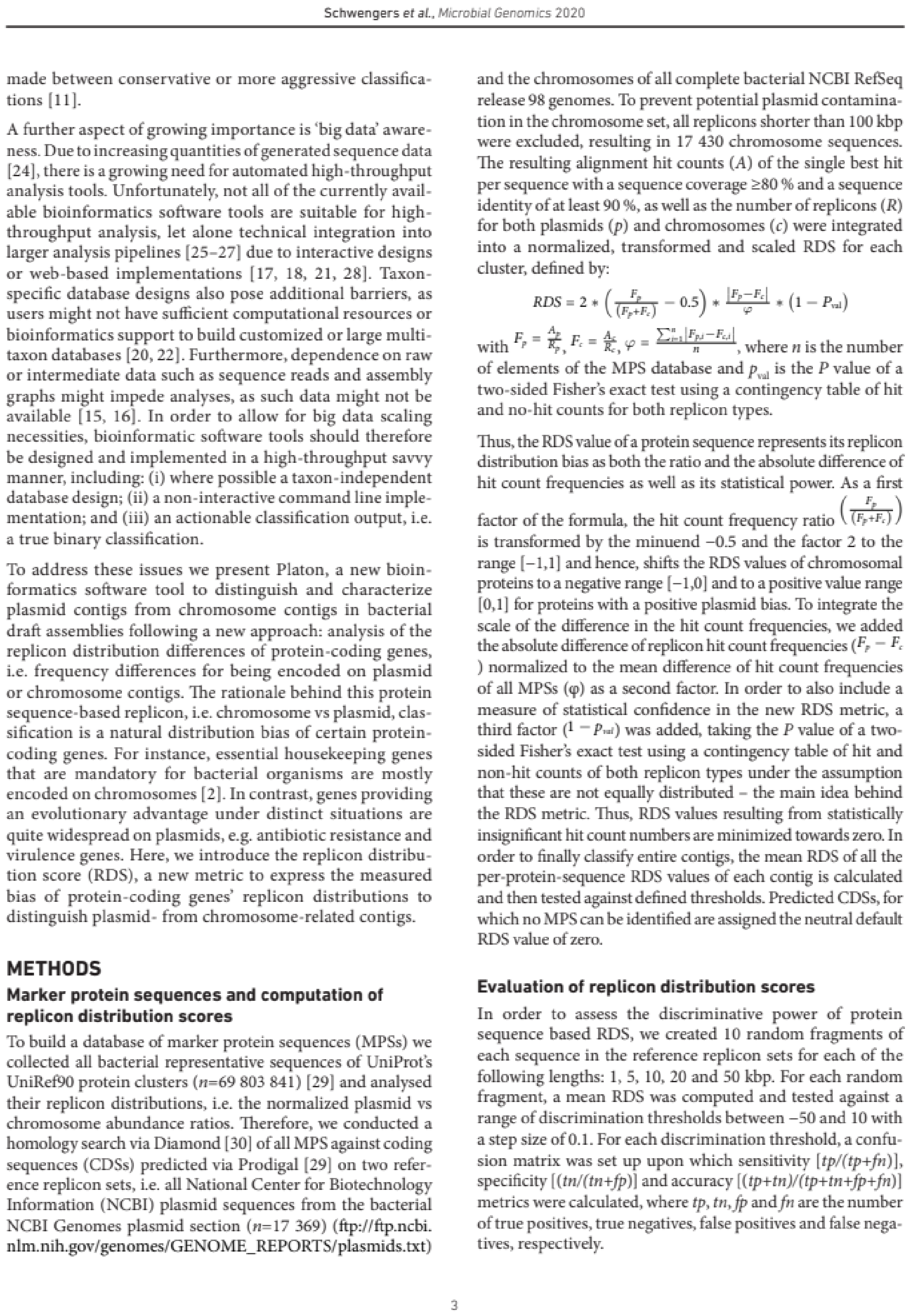
Such analysis can be performed by plasmid DNA isolation followed by sequencing [9]. However, due to decreased sequencing costs, it is now affordable and often easier to sequence the entire genome of bacterial organisms using next-generation whole-genome shotgun sequencing [10]. Furthermore, this approach allows the reanalysis of already sequenced genomes to identify plasmids that have not been detected before. Unfortunately, this introduces a new issue that needs to be addressed: plasmid and chromosomal contigs are mixed in draft assemblies and need to be distinguished from each other.

This task, however, is hard to achieve for biological and technical reasons [11]. Plasmids often contain mobile genetic elements, e.g. transposons and integrons, which are drivers for the genetic exchange between different DNA replicons and regions [12, 13]. Hence, these genetic elements are often encoded on both replicon types and thus the origin of DNA fragments encoding such elements is hard to predict. Modern short-read assemblers add additional intricacy, further aggravating these issues, as they are notoriously hard pressed to correctly assemble repetitive regions such as the aforementioned DNA elements [14]. To tackle this issue, many new bioinformatic tools have recently been developed, following different approaches: (i) Recycler and plasmidSPAdes [15, 16] exploit coverage variations of sequenced DNA fragments within a genome; (ii) PLACNET investigates paired-end reads linking contig ends [17]; (iii) PlasmidFinder searches for plasmid specific motifs, i.e. incompatibility groups [18]; (iv) cBar, PlasFlow and mlPlasmids use machine learning methods to classify k-mer frequencies [19–21]; (v) PlaScope and PlasmidSeeker perform fast k-mer-based database searches of known plasmid sequences [22, 23]; Recycler additionally exploits information on circularization [15]. Overall, each approach provides unique advantages and drawbacks. For example, approaches based on sequencing coverage variations are unable to detect plasmids with copy numbers equal to the chromosome, whereas sequence motif- and k-mer-based methods tend to identify only known plasmids. This often leads to distinct profiles in terms of sensitivity and specificity, which are often biased towards one of the metrics, and as this impacts on the conducted analysis a choice must be

Impact Statement

Plasmids play a vital role in the spread of antibiotic resistance and pathogenicity genes. The increasing numbers of clinical outbreaks involving resistant pathogens worldwide pushed the scientific community to increase their efforts to comprehensively investigate bacterial genomes. Due to the maturation of next-generation sequencing technologies, entire bacterial genomes, including plasmids, are now sequenced on a huge scale. To analyse draft assemblies, a mandatory first step is to separate plasmid from chromosome contigs. Recently, many bioinformatic tools have emerged to tackle this issue. Unfortunately, several tools are only implemented as interactive or web-based tools, making them unavailable for the necessary high-throughput analysis of large datasets. Other tools providing such a high-throughput implementation, however, often come with certain drawbacks, e.g. providing taxon-specific databases only, not providing actionable, i.e. true, binary classification, or showing classification performance that is biased towards either sensitivity or specificity. Here, we introduce the tool Platon, implementing a new replicon distribution-based approach combined with higher-level contig characterizations to address the aforementioned issues. In addition to the plasmid detection within draft assemblies, Platon provides the user with valuable information on certain higher-level contig characterizations. We show that Platon provides a balanced classification performance as well as a scalable implementation for high-throughput analyses. We therefore consider Platon to be a powerful, species-independent and flexible tool to scan large quantities of bacterial whole-genome sequencing data for their plasmid content.

7.3 Platon: identification and characterization of bacterial plasmid contigs from short-read draft assemblies exploiting protein-sequence-based replicon distribution scores.



made between conservative or more aggressive classifications [11].

A further aspect of growing importance is 'big data' awareness. Due to increasing quantities of generated sequence data [24], there is a growing need for automated high-throughput analysis tools. Unfortunately, not all of the currently available bioinformatics software tools are suitable for high-throughput analysis, let alone technical integration into larger analysis pipelines [25–27] due to interactive designs or web-based implementations [17, 18, 21, 28]. Taxon-specific database designs also pose additional barriers, as users might not have sufficient computational resources or bioinformatics support to build customized or large multi-taxon databases [20, 22]. Furthermore, dependence on raw or intermediate data such as sequence reads and assembly graphs might impede analyses, as such data might not be available [15, 16]. In order to allow for big data scaling necessities, bioinformatic software tools should therefore be designed and implemented in a high-throughput savvy manner, including: (i) where possible a taxon-independent database design; (ii) a non-interactive command line implementation; and (iii) an actionable classification output, i.e. a true binary classification.

To address these issues we present Platon, a new bioinformatics software tool to distinguish and characterize plasmid contigs from chromosome contigs in bacterial draft assemblies following a new approach: analysis of the replicon distribution differences of protein-coding genes, i.e. frequency differences for being encoded on plasmid or chromosome contigs. The rationale behind this protein sequence-based replicon, i.e. chromosome vs plasmid, classification is a natural distribution bias of certain protein-coding genes. For instance, essential housekeeping genes that are mandatory for bacterial organisms are mostly encoded on chromosomes [2]. In contrast, genes providing an evolutionary advantage under distinct situations are quite widespread on plasmids, e.g. antibiotic resistance and virulence genes. Here, we introduce the replicon distribution score (RDS), a new metric to express the measured bias of protein-coding genes' replicon distributions to distinguish plasmid- from chromosome-related contigs.

METHODS

Marker protein sequences and computation of replicon distribution scores

To build a database of marker protein sequences (MPSs) we collected all bacterial representative sequences of UniProt's UniRef90 protein clusters (n=69 803 841) [29] and analysed their replicon distributions, i.e. the normalized plasmid vs chromosome abundance ratios. Therefore, we conducted a homology search via Diamond [30] of all MPS against coding sequences (CDSs) predicted via Prodigal [29] on two reference replicon sets, i.e. all National Center for Biotechnology Information (NCBI) plasmid sequences from the bacterial NCBI Genomes plasmid section (n=17 369) (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/plasmids.txt) and the chromosomes of all complete bacterial NCBI RefSeq release 98 genomes. To prevent potential plasmid contamination in the chromosome set, all replicons shorter than 100 kbp were excluded, resulting in 17 430 chromosome sequences. The resulting alignment hit counts (A) of the single best hit per sequence with a sequence coverage ≥80 % and a sequence identity of at least 90 %, as well as the number of replicons (R) for both plasmids (p) and chromosomes (c) were integrated into a normalized, transformed and scaled RDS for each cluster, defined by:

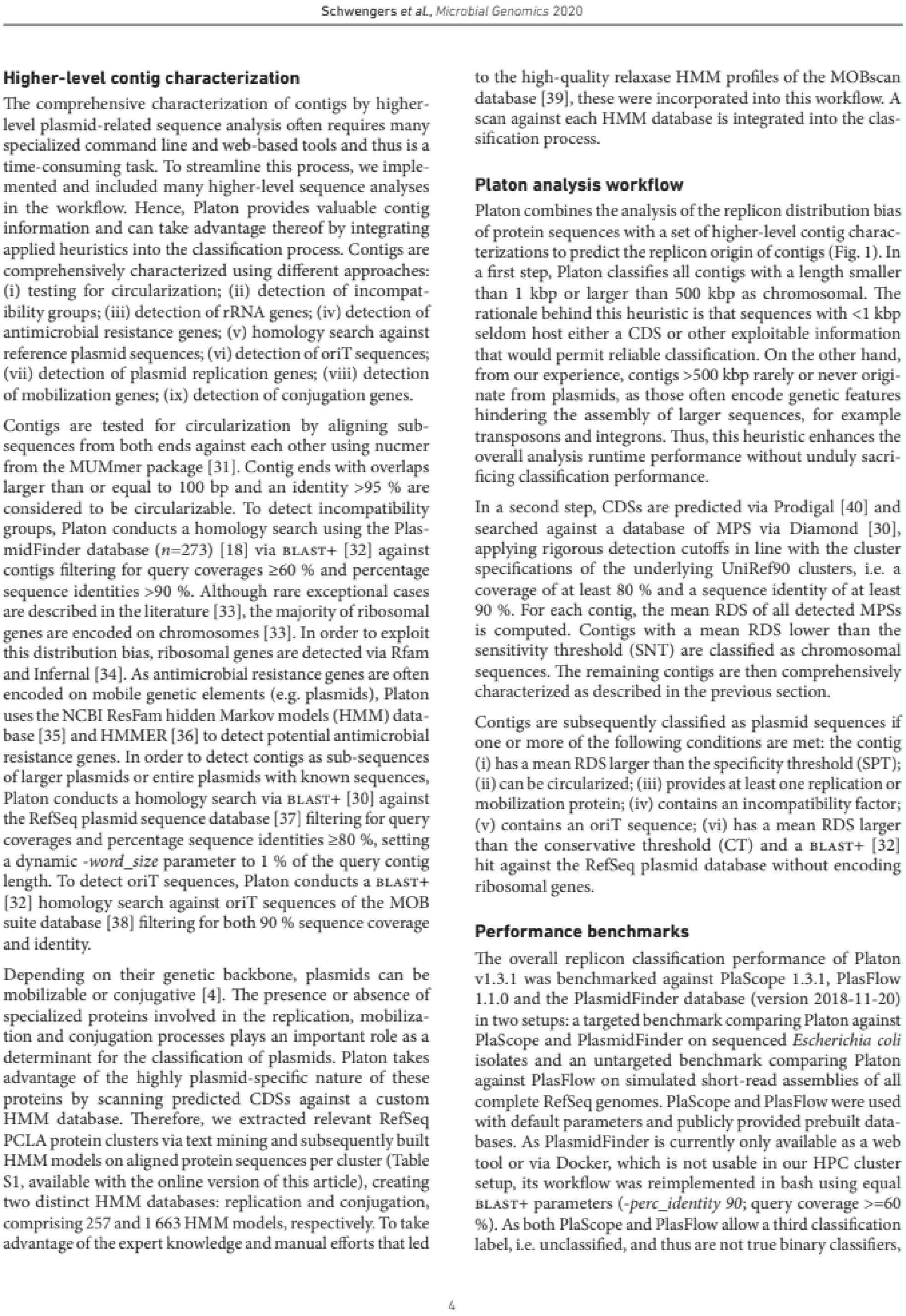
$$RDS = 2 * \left(\frac{F_p}{(F_p+F_c)} - 0.5\right) * \frac{|F_p - F_c|}{\varphi} * (1 - P_{\text{val}})$$

with $F_p = \frac{A_p}{R_p}$, $F_c = \frac{A_c}{R_c}$, $\varphi = \frac{\sum_{i=1}^{n} |F_{p,i} - F_{c,i}|}{n}$, where n is the number of elements of the MPS database and p_{val} is the P value of a two-sided Fisher's exact test using a contingency table of hit and no-hit counts for both replicon types.

Thus, the RDS value of a protein sequence represents its replicon distribution bias as both the ratio and the absolute difference of hit count frequencies as well as its statistical power. As a first factor of the formula, the hit count frequency ratio $\left(\frac{F_p}{(F_p+F_c)}\right)$ is transformed by the minuend −0.5 and the factor 2 to the range [−1,1] and hence, shifts the RDS values of chromosomal proteins to a negative range [−1,0] and to a positive value range [0,1] for proteins with a positive plasmid bias. To integrate the scale of the difference in the hit count frequencies, we added the absolute difference of replicon hit count frequencies ($F_p - F_c$) normalized to the mean difference of hit count frequencies of all MPSs (φ) as a second factor. In order to also include a measure of statistical confidence in the new RDS metric, a third factor ($1 - P_{val}$) was added, taking the P value of a two-sided Fisher's exact test using a contingency table of hit and non-hit counts of both replicon types under the assumption that these are not equally distributed – the main idea behind the RDS metric. Thus, RDS values resulting from statistically insignificant hit count numbers are minimized towards zero. In order to finally classify entire contigs, the mean RDS of all the per-protein-sequence RDS values of each contig is calculated and then tested against defined thresholds. Predicted CDSs, for which no MPS can be identified are assigned the neutral default RDS value of zero.

Evaluation of replicon distribution scores

In order to assess the discriminative power of protein sequence based RDS, we created 10 random fragments of each sequence in the reference replicon sets for each of the following lengths: 1, 5, 10, 20 and 50 kbp. For each random fragment, a mean RDS was computed and tested against a range of discrimination thresholds between −50 and 10 with a step size of 0.1. For each discrimination threshold, a confusion matrix was set up upon which sensitivity [$tp/(tp+fn)$], specificity [$(tn/(tn+fp))$] and accuracy [$(tp+tn)/(tp+tn+fp+fn)$] metrics were calculated, where tp, tn, fp and fn are the number of true positives, true negatives, false positives and false negatives, respectively.



Higher-level contig characterization

The comprehensive characterization of contigs by higher-level plasmid-related sequence analysis often requires many specialized command line and web-based tools and thus is a time-consuming task. To streamline this process, we implemented and included many higher-level sequence analyses in the workflow. Hence, Platon provides valuable contig information and can take advantage thereof by integrating applied heuristics into the classification process. Contigs are comprehensively characterized using different approaches: (i) testing for circularization; (ii) detection of incompatibility groups; (iii) detection of rRNA genes; (iv) detection of antimicrobial resistance genes; (v) homology search against reference plasmid sequences; (vi) detection of oriT sequences; (vii) detection of plasmid replication genes; (viii) detection of mobilization genes; (ix) detection of conjugation genes.

Contigs are tested for circularization by aligning sub-sequences from both ends against each other using nucmer from the MUMmer package [31]. Contig ends with overlaps larger than or equal to 100 bp and an identity >95 % are considered to be circularizable. To detect incompatibility groups, Platon conducts a homology search using the PlasmidFinder database (*n*=273) [18] via BLAST+ [32] against contigs filtering for query coverages ≥60 % and percentage sequence identities >90 %. Although rare exceptional cases are described in the literature [33], the majority of ribosomal genes are encoded on chromosomes [33]. In order to exploit this distribution bias, ribosomal genes are detected via Rfam and Infernal [34]. As antimicrobial resistance genes are often encoded on mobile genetic elements (e.g. plasmids), Platon uses the NCBI ResFam hidden Markov models (HMM) database [35] and HMMER [36] to detect potential antimicrobial resistance genes. In order to detect contigs as sub-sequences of larger plasmids or entire plasmids with known sequences, Platon conducts a homology search via BLAST+ [30] against the RefSeq plasmid sequence database [37] filtering for query coverages and percentage sequence identities ≥80 %, setting a dynamic *-word_size* parameter to 1 % of the query contig length. To detect oriT sequences, Platon conducts a BLAST+ [32] homology search against oriT sequences of the MOB suite database [38] filtering for both 90 % sequence coverage and identity.

Depending on their genetic backbone, plasmids can be mobilizable or conjugative [4]. The presence or absence of specialized proteins involved in the replication, mobilization and conjugation processes plays an important role as a determinant for the classification of plasmids. Platon takes advantage of the highly plasmid-specific nature of these proteins by scanning predicted CDSs against a custom HMM database. Therefore, we extracted relevant RefSeq PCLA protein clusters via text mining and subsequently built HMM models on aligned protein sequences per cluster (Table S1, available with the online version of this article), creating two distinct HMM databases: replication and conjugation, comprising 257 and 1 663 HMM models, respectively. To take advantage of the expert knowledge and manual efforts that led to the high-quality relaxase HMM profiles of the MOBscan database [39], these were incorporated into this workflow. A scan against each HMM database is integrated into the classification process.

Platon analysis workflow

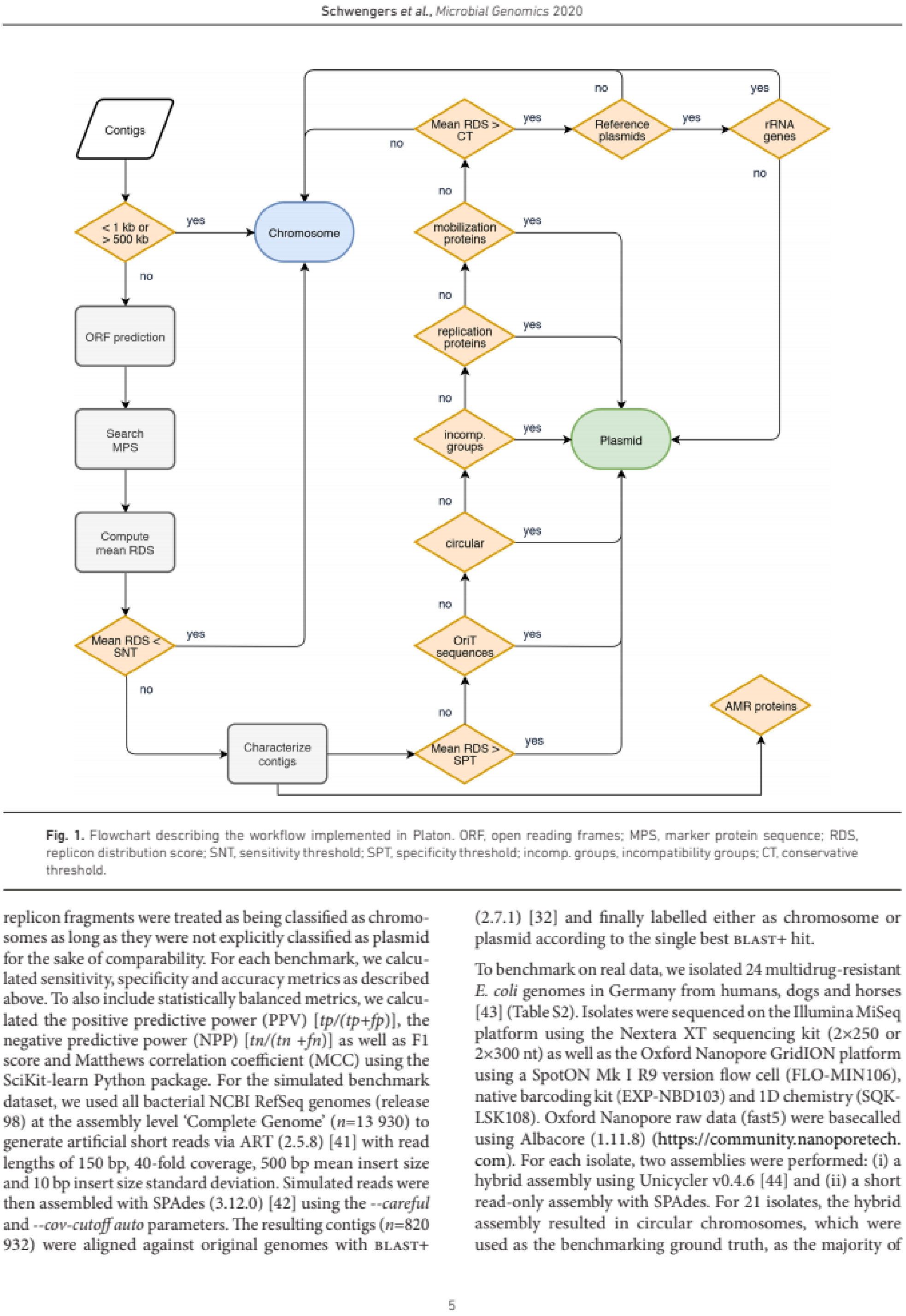
Platon combines the analysis of the replicon distribution bias of protein sequences with a set of higher-level contig characterizations to predict the replicon origin of contigs (Fig. 1). In a first step, Platon classifies all contigs with a length smaller than 1 kbp or larger than 500 kbp as chromosomal. The rationale behind this heuristic is that sequences with <1 kbp seldom host either a CDS or other exploitable information that would permit reliable classification. On the other hand, from our experience, contigs >500 kbp rarely or never originate from plasmids, as those often encode genetic features hindering the assembly of larger sequences, for example transposons and integrons. Thus, this heuristic enhances the overall analysis runtime performance without unduly sacrificing classification performance.

In a second step, CDSs are predicted via Prodigal [40] and searched against a database of MPS via Diamond [30], applying rigorous detection cutoffs in line with the cluster specifications of the underlying UniRef90 clusters, i.e. a coverage of at least 80 % and a sequence identity of at least 90 %. For each contig, the mean RDS of all detected MPSs is computed. Contigs with a mean RDS lower than the sensitivity threshold (SNT) are classified as chromosomal sequences. The remaining contigs are then comprehensively characterized as described in the previous section.

Contigs are subsequently classified as plasmid sequences if one or more of the following conditions are met: the contig (i) has a mean RDS larger than the specificity threshold (SPT); (ii) can be circularized; (iii) provides at least one replication or mobilization protein; (iv) contains an incompatibility factor; (v) contains an oriT sequence; (vi) has a mean RDS larger than the conservative threshold (CT) and a BLAST+ [32] hit against the RefSeq plasmid database without encoding ribosomal genes.

Performance benchmarks

The overall replicon classification performance of Platon v1.3.1 was benchmarked against PlaScope 1.3.1, PlasFlow 1.1.0 and the PlasmidFinder database (version 2018-11-20) in two setups: a targeted benchmark comparing Platon against PlaScope and PlasmidFinder on sequenced *Escherichia coli* isolates and an untargeted benchmark comparing Platon against PlasFlow on simulated short-read assemblies of all complete RefSeq genomes. PlaScope and PlasFlow were used with default parameters and publicly provided prebuilt databases. As PlasmidFinder is currently only available as a web tool or via Docker, which is not usable in our HPC cluster setup, its workflow was reimplemented in bash using equal BLAST+ parameters (*-perc_identity 90*; query coverage >=60 %). As both PlaScope and PlasFlow allow a third classification label, i.e. unclassified, and thus are not true binary classifiers,



Contigs
< 1 kb or > 500 kb
yes
no
Chromosome
ORF prediction
Search MPS
Compute mean RDS
Mean RDS < SNT
yes
no
Characterize contigs
Mean RDS > SPT
yes
no
OriT sequences
yes
no
circular
yes
no
incomp. groups
yes
no
replication proteins
yes
no
mobilization proteins
yes
no
Mean RDS > CT
yes
no
Reference plasmids
yes
no
rRNA genes
yes
no
Plasmid
AMR proteins

Fig. 1. Flowchart describing the workflow implemented in Platon. ORF, open reading frames; MPS, marker protein sequence; RDS, replicon distribution score; SNT, sensitivity threshold; SPT, specificity threshold; incomp. groups, incompatibility groups; CT, conservative threshold.

replicon fragments were treated as being classified as chromosomes as long as they were not explicitly classified as plasmid for the sake of comparability. For each benchmark, we calculated sensitivity, specificity and accuracy metrics as described above. To also include statistically balanced metrics, we calculated the positive predictive power (PPV) [*tp/*(*tp*+*fp*)], the negative predictive power (NPP) [*tn*/(*tn* +*fn*)] as well as F1 score and Matthews correlation coefficient (MCC) using the SciKit-learn Python package. For the simulated benchmark dataset, we used all bacterial NCBI RefSeq genomes (release 98) at the assembly level 'Complete Genome' (*n*=13 930) to generate artificial short reads via ART (2.5.8) [41] with read lengths of 150 bp, 40-fold coverage, 500 bp mean insert size and 10 bp insert size standard deviation. Simulated reads were then assembled with SPAdes (3.12.0) [42] using the *--careful* and *--cov-cutoff auto* parameters. The resulting contigs (*n*=820 932) were aligned against original genomes with BLAST+ (2.7.1) [32] and finally labelled either as chromosome or plasmid according to the single best BLAST+ hit.

To benchmark on real data, we isolated 24 multidrug-resistant *E. coli* genomes in Germany from humans, dogs and horses [43] (Table S2). Isolates were sequenced on the Illumina MiSeq platform using the Nextera XT sequencing kit (2×250 or 2×300 nt) as well as the Oxford Nanopore GridION platform using a SpotON Mk I R9 version flow cell (FLO-MIN106), native barcoding kit (EXP-NBD103) and 1D chemistry (SQK-LSK108). Oxford Nanopore raw data (fast5) were basecalled using Albacore (1.11.8) (https://community.nanoporetech.com). For each isolate, two assemblies were performed: (i) a hybrid assembly using Unicycler v0.4.6 [44] and (ii) a short read-only assembly with SPAdes. For 21 isolates, the hybrid assembly resulted in circular chromosomes, which were used as the benchmarking ground truth, as the majority of remaining contigs thus originate from unclosed plasmids. The remaining three isolates with unclosed chromosomes were excluded from the benchmark dataset, as the former requirement was not fulfilled. Short-read contigs <1 kbp were discarded. The remaining contigs (*n*=1 337) were then aligned against closed hybrid assemblies as described above. The raw sequencing data for all 24 isolates are available as NCBI BioProjects (PRJNA505407, PRJNA387731).

RESULTS AND DISCUSSION

Creation of the MPS database and RDS-based inference of contig origins

The proposed new metric RDS exploits the natural distribution biases of protein-coding genes between chromosomes and plasmids to classify the origin of contigs from short-read assemblies. In order to investigate and test this rationale, we aligned a broad range of bacterial protein sequences (*n*=69 803 841) from UniProt's UniRef90 protein cluster representative sequences against a set of known chromosome and plasmid reference replicons from the NCBI RefSeq and NCBI Genomes databases and 12 795 544 of these protein sequences could be aligned to at least 1 replicon. For each of these protein sequences, a two-sided Fisher's exact test was conducted and sequences with a *P* value of 1 were excluded. The remaining protein sequences (*n*=4 108 727), along with their RDS values, product description and sequence lengths, were then used to compile the final MPS database. For 99.5 % of these protein sequences (*n*=4 089 068) a transformed hit count ratio smaller than −0.5 (*n*=3 600 927) or larger than 0.5 (*n*=488 141) was computed, indicating a rather unequal distribution between chromosomes and plasmids (Fig. 2). However, only a minor fraction of 7.8 % (*n*=322 151) of all MPSs had a normalized alignment hit count sum regarding both replicon types larger than 0.001. Hence, the majority of MPS database sequences were relatively rarely detected on average. These findings endorse the incorporation of the statistical significance of each MPS replicon distribution as well as the scaling by the absolute difference of replicon hit count frequencies in order to raise the contribution of abundant protein sequences and decrease the contribution of rare protein sequences, for which insufficient data are available in the reference replicon sets.

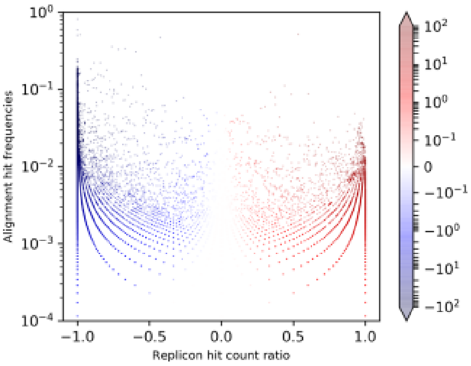


Fig. 2. Replicon distribution and alignment hit frequencies of marker protein sequences. Shown here are summed plasmid and chromosome alignment hit frequencies per marker protein sequence plotted against plasmid/chromosome hit count ratios scaled to [−1, 1]; Hue: normalized replicon distribution score values (min=−100, max=100), hit count outliers below 10^{-4} and above 1 are discarded for the sake of readability.

In order to assess the discriminative performance of RDS regarding the replicon origin of contigs, we tested a broad range of thresholds computing sensitivity, specificity and accuracy metrics. The sensitivity, specificity and accuracy values for a range of RDS thresholds are plotted in Fig. 3. The sensitivity and specificity curves follow a sharp inflection point near the default RDS value, i.e. 0. We attribute this behaviour to contigs harbouring protein sequences that are not covered by the MPS database. To overcome this limitation and achieve both sensitive and specific classifications, we defined three distinct thresholds: (i) an SNT; (ii) an SPT; (iii) a CT set to 95 % sensitivity, 99.9 % specificity and the highest accuracy, respectively. Thus, contigs with an RDS smaller than the SNT can be classified as chromosomal while still retaining 95 % of all plasmid contigs. Correspondingly, contigs with an RDS larger than the SPT can be classified as plasmid fragments achieving a specificity ≥99.9 %. To compute actual values for these thresholds, we conducted classifications of Monte Carlo replicon fragment simulations (*n*=1 564 639) by which the following values were established: SNT=−7.7, SPT=0.4 and CT=0.1 at a maximal accuracy of 84.1 %. These values surround the inflection point near 0 and were henceforth used as the final discrimination thresholds in the Platon implementation.

To finally assess the RDS-based contig classification, a comprehensive performance benchmark was conducted. To do this, we created simulated short reads based on all complete NCBI RefSeq genomes (*n*=13 930) covering a broad range of bacterial taxa. The resulting short reads were then reassembled into contigs (*n*=820 392), which were aligned back to the original genomes, thus creating our ground truth. This benchmark dataset comprised a total of 63 107 true plasmid contigs. All contigs were classified by their mean RDS value, applying the computed SNT and SPT thresholds. This RDS workflow classified 38 197 plasmid contigs and 754 082 chromosomal contigs correctly, thus achieving an accuracy of 0.966 and a sensitivity of 0.605, as well as an F1 score of 0.731 and an MCC of 0.732, calculated using the following confusion matrix: *tp*=38 197, *tn*=754 082, *fp*=3 203, *fn*=24 910.

Although the RDS approach achieved an accuracy of 0.966, it still misclassified 24 910 true plasmid contigs and 3 203 true chromosomal contigs. It is common knowledge that certain proteins are encoded on both replicon types, for instance, relaxases and type4-coupling proteins (T4CP) – key proteins of integrative conjugative elements [45]. To assess the discriminative power of the RDS metric on these widespread

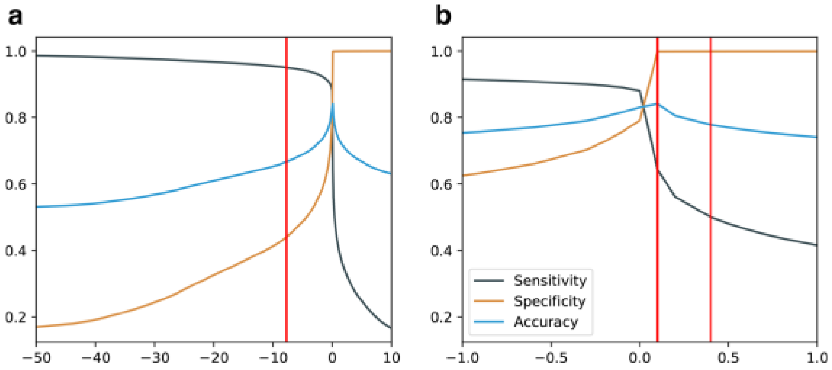


Fig. 3. Evaluation statistics for replicon distribution score thresholds. Sensitivity, specificity and accuracy values are plotted against replicon distribution score threshold ranges. (a) Overview threshold range [−50,10]. (b) Detailed threshold range [−1,1]. Sensitivity is in black, specificity is in brown and accuracy is in blue. Red vertical lines from left to right: sensitivity threshold (−7.7), conservative threshold (0.1) and specificity threshold (0.4).

protein classes we extracted a set of 4 683 relaxase and 2 151 T4CP clusters from the MPS database by MOBscan [39] and TXSscan [46] HMM profile searches and investigated the range of related RDS values (Fig. S1); 73 and 66 % from the relaxase (*n*=3 321) and T4CP (*n*=1 436) protein clusters had an RDS between −0.5 and 0.5 and thus can be considered to be quite equally distributed. Small contigs solely or mainly encoding these protein sequences could therefore be especially hard to classify by the RDS metric. However, we also found 817 and 411 protein clusters that were quite chromosomally biased with a related RDS below −0.5 and extremes reaching values of −64.96 and −37.47 for the relaxases and T4CP, respectively. In addition, 445 and 304 protein clusters were quite plasmid biased with a related RDS above 0.5 and extremes reaching values of as high as 109.60 and 79.76 for the relaxases and T4CP, respectively. The latter protein clusters constitute approximately a quarter and a third of all relaxase and T4CP MPS subsets and have highly discriminative related RDS values. Hence, although there are protein classes harbouring many fairly equally distributed protein clusters, e.g. the analysed relaxases and T4CP, which are often encoded in very-hard-to-classify integrative conjugative elements, we still found MPSs with a strong predictive power regarding the replicon origin of a contig.

Performance of the entire Platon workflow

As shown in the simulated short-read benchmark, the RDS metric achieved a high accuracy (ac=0.966) but rather moderate sensitivity (sn=0.605) due to the high number of false negatives (fn=24 910). In order to increase the detection rate of true plasmid contigs, the Platon workflow additionally comprises higher-level plasmid-related contig characterizations that serve as a basis for several heuristics. As both the protein homology search and the contig characterizations of large plasmids are computationally expensive, contigs >500 kbp are automatically assigned to the chromosome. To assess the potentially negative impact of this heuristic on the classification performance, contig length distributions for both replicon types within the simulated short-read dataset (Fig. S2) were investigated. In line with the smaller plasmid contig length on average, only 119 of 63 107 plasmid contigs were actually larger than 500 kbp compared to 15 750 of 757 285 chromosome contigs. Hence, only 0.19 % of all plasmid contigs were erroneously assigned to the chromosome, but 99.25 % of all contigs larger than 500 kbp were correctly classified by this heuristic, which thus qualifies as an eligible tradeoff between sensitivity and runtime.

To measure and assess the overall classification performance of the entire implemented workflow (Fig. 1), we conducted two benchmarks against contemporary command line tools: an untargeted benchmark against PlasFlow on the aforementioned simulated short-read dataset as well as a targeted benchmark against PlaScope and PlasmidFinder on sequenced *E. coli* isolates.

Performance benchmark on taxonomically diverse simulated short-read assemblies

To assess the performance of the extended Platon workflow in an untargeted, i.e. taxon- independent, setup, we conducted a comprehensive benchmark against PlasFlow, a contemporary plasmid prediction tool for metagenomics that was presented to also be eligible for the recruitment of plasmid contigs from isolates. For this benchmark, all complete bacterial NCBI RefSeq genomes (*n*=13 930) covering a broad range of bacterial taxa were used to simulate short reads that were *de novo* assembled. The resulting contigs were then aligned back onto original genomes. A confusion matrix as well as common

Table 1. Performance benchmark results computed contig-wise on simulated short-read data

Metric	PlasFlow	Platon
Accuracy	0.871	**0.976**
Sensitivity	0.729	**0.766**
Specificity	0.883	**0.993**
PPV	0.341	**0.902**
NPV	0.975	**0.981**
F1	0.465	**0.828**
MCC	0.440	**0.818**
TP	45 999	**48 333**
TN	668 573	**752 080**
FP	88 712	**5 277**
FN	17 108	**14 774**

classifier performance metrics aggregated for all contigs (*n*=820 392) are shown in Table 1. In this benchmark Platon recruited 48 333 and PlasFlow 45 999 true plasmid contigs, resulting in comparable sensitivity and negative predictive values (NPV) of 0.762 and 0.729 and 0.98 and 0.975, respectively. However, PlasFlow predicted 17 times more false positives (fp=88 712) than Platon (fp=5 277). Due to the notably lower number of false positives, Platon clearly outperformed PlasFlow in terms of accuracy, specificity and positive predictive value (PPV), as well as the balanced metrics F1 score and Matthew's correlation coefficient (MCC). An overview of how many contigs could be classified by which RDS threshold and heuristic filter is given in Table S3.

Due to different contig lengths, the mere number of correctly classified contigs might not always be congruent with the recruited plasmid content, which could play a vital role in downstream analyses, e.g. the recruitment of plasmid-borne genes or sequence motifs, such as oriT and oriV. Hence, benchmarks that only measure the number of classified contigs might, to some extent, be misleading, and so we complemented the former benchmark with a genomic content-based view calculating an additional confusion matrix based on classified DNA nucleotides (Table S4). Fig. 4 provides a combined view on both benchmark setups. In this complementary benchmark, the specificity values for PlasFlow increased from 0.883 contig-wise to 0.979 nucleotide-wise compared to stable and higher values for Platon (contig-wise=0.993; nucleotide-wise=0.995). The accuracy values also increased from 0.871 contig-wise to 0.974 nucleotide-wise for PlasFlow, whereas the accuracy values achieved by Platon only improved slightly (contig-wise=0.976; nucleotide-wise=0.99). Taking into account the genomic content of classified contigs revealed a performance improvement of PlasFlow in terms of accuracy and specificity, but it still fell slightly below Platon. However, PlasFlow predicted 4.3 times more false-positive plasmid nucleotides (fp=1 115.3 mbp) than Platon (fp=260.9 mbp), in line with the contig-wise benchmark.

The taxonomic compositions of training datasets for machine learning approaches and prebuilt databases can have a severe impact on benchmark performance and the results of analyses. To assess a potential bias towards certain taxa we additionally analysed the taxonomic distribution of the recruited plasmid contigs of the simulated short-read dataset binned to the genus level (Fig. 5). The underlying benchmark dataset contained true plasmid contigs from 469 distinct genera and 1234 species. From these, Platon recruited plasmid contigs from 434 genera, whereas PlasFlow recruited plasmid contigs from 384 genera (Table S5). For both tools, the three taxa *Escherichia*, *Klebsiella* and *Enterococcus* accounted for nearly 40 % of the recruited sequences alike the taxonomic profile of the underlying benchmark dataset in which the aforementioned

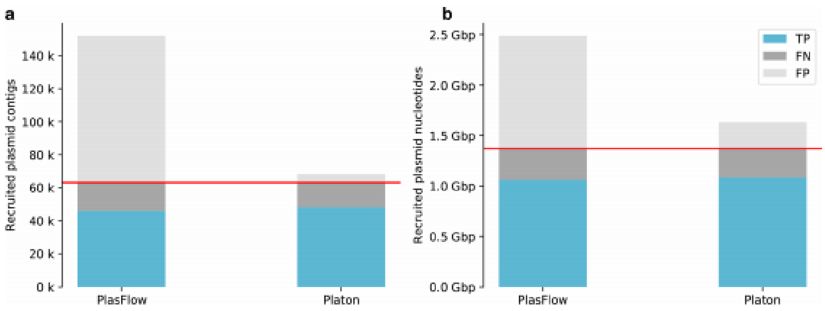


Fig. 4. Performance benchmark metrics on simulated short-read data. A performance benchmark was conducted on all complete bacterial genomes of the NCBI RefSeq database, assembling simulated short reads and subsequently realigning them onto original genomes. For scaling reasons and the sake of readability, true negatives were discarded. (a) Benchmark results calculated contig-wise. Horizontal red line, total number of true plasmid contigs. (b) Benchmark results calculated nucleotide-wise. Horizontal red line, total number of true plasmid DNA nucleotides.

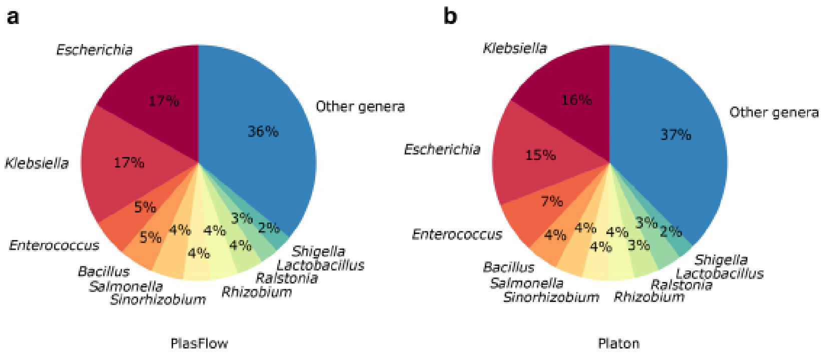


Fig. 5. Taxonomic distribution of recruited plasmid contigs. The taxonomic distribution of the recruited plasmid contigs for the simulated benchmark dataset is shown binned to the genus level. Taxa accounting for less than 2 % are grouped as 'others'. (a) PlasFlow; (b) Platon.

taxa accounted for 26 %. On a species level, Platon and PlasFlow recruited plasmid contigs from 1 128 and 1 014 distinct species, respectively, in line with the aforementioned genus-level results. Although PlasFlow was developed as an untargeted tool for metagenomics, Platon recruited plasmid contigs from a wider taxonomic range, thus demonstrating the competitive edge of the taxon-independent RDS approach complemented by contig characterization heuristics.

Targeted performance benchmark on sequenced *E. coli* isolates

Simulated data seldom reflect the existing biological and technical complexity and the plethora of potential pitfalls. Hence, we additionally benchmarked the Platon workflow on real data in a targeted setup. We compared the performance of Platon against PlaScope and PlasmidFinder, which were both published as targeted approaches for the plasmid prediction within whole-genome sequencing data. PlaScope provides a precompiled *E. coli* database for download, which was used in this benchmark, and PlasmidFinder was specifically designed for the analysis of *Enterobacteriaceae* genomes. As the PlasmidFinder database is part of Platon's contig characterization, we assessed its performance to transparently compare both tools side by side. For this benchmark the genomes of 24 *E. coli* isolates were sequenced using both Illumina short-read and Oxford Nanopore long-read technologies. For 21 isolates the hybrid assemblies resulted in closed chromosomes, which were used as the ground truth data. Contigs from short read-only assemblies (*n*=1 337) were aligned to the closed assemblies and used as the actual benchmark data. Table 2 shows the confusion matrix as well as computed benchmark metrics. PlasmidFinder achieved the lowest false-positive rate (fp=14) resulting in the highest specificity of 0.987, closely followed by Platon (*sp*=0.966) and PlaScope (*sp*=0.952), but showed the lowest true-positive rate (tp=57) and sensitivity (sn=0.223), thus performing worse than Platon (sn=0.699) and PlaScope (sn=0.684). With regard to accuracy, PPV, NPV, F1 score and MCC metrics, Platon and PlaScope performed nearly on par, although Platon was slightly ahead on each. Both tools performed better than PlasmidFinder on these metrics. This was especially true for the balanced metrics F1 score and MCC, for which Platon and Plascope clearly outperformed PlasmidFinder.

Similarly, with the simulated short-read benchmark we also compared the performances of Platon, PlaScope and PlasmidFinder, taking into account the amount of genomic content (Fig. 6) computed on a nucleotide-wise confusion matrix (Table S6). The nucleotide-wise results were in line with those calculated contig-wise: PlasmidFinder had the lowest number of false positives, but also detected remarkably fewer plasmid nucleotides than PlaScope and Platon. The latter two detected a nearly equal quantity of plasmid

Table 2. Performance benchmark results contig-wise on sequenced isolate short-read data

Metric	PlaScope	PlasmidFinder	Platon
Accuracy	0.901	0.841	**0.915**
Sensitivity	0.684	0.223	**0.699**
Specificity	0.952	**0.987**	0.966
PPV	0.771	0.803	**0.829**
NPV	0.927	0.843	**0.931**
F1	0.725	0.349	**0.758**
MCC	0.666	0.368	**0.711**
TP	175	57	**179**
TN	1 029	**1 067**	1 044
FP	52	**14**	37
FN	81	199	**77**

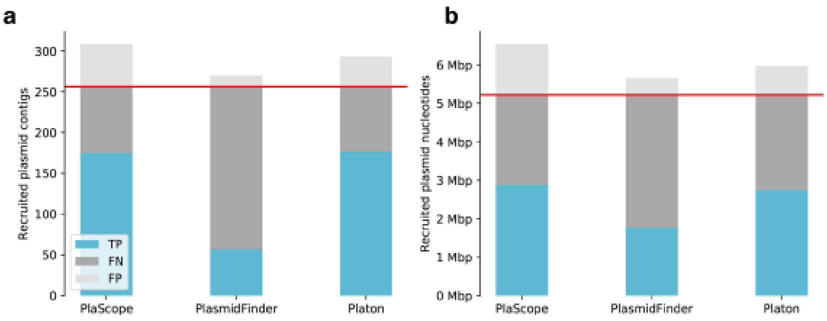


Fig. 6. Performance benchmark metrics on real short-read data. A performance benchmark was conducted on 21 *E. coli* genomes, for which both short-read draft assemblies and complete genomes via hybrid assemblies were available. For scaling reasons and the sake of readability, true negatives were discarded. (a) Benchmark results calculated contig-wise. Horizontal red line, total number of true plasmid contigs. (b) Benchmark results calculated nucleotide-wise. Horizontal red line, total number of true plasmid DNA nucleotides.

content, with Platon predicting notably fewer false positives than PlaScope.

Conclusion

Due to the complex nature of plasmid fragments, replicon type classification, i.e. prediction of origin, for contigs resulting from short-read draft assemblies is a difficult task. Many different methods and tools have recently been described in the literature, but few work on draft assemblies only, are implemented in a high-throughput savvy manner or provide statistically balanced predictions in an untargeted, i.e. taxon-independent manner.

To tackle this issue, we investigated the natural distribution biases of protein-coding genes between chromosomes and plasmids for a large set of protein sequences in bacteria. In this study, we defined, computed and tested statistical discrimination thresholds for the introduced new metric RDS and showed that it is a feasible approach to the problem. However, small contigs without sufficient protein sequences or contigs encoding for protein sequences that were either not covered by the MPS database or equally distributed between chromosomes and plasmids remained hard to classify correctly. However, even for the protein classes relaxases and T4CP, which are often found on notoriously hard-to-classify integrative conjugative elements, we found protein sequences with strong predictive power. To mitigate these drawbacks and improve the overall sensitivity, we complemented this approach with several heuristics exploiting higher-level plasmid-related sequence characterizations. We implemented this new workflow in a software tool called Platon and conducted benchmarks against three contemporary software tools, i.e. PlaScope, PlasFlow and PlasmidFinder on both simulated short-read data and sequenced isolates.

Analysing a large set of diverse bacterial species, Platon achieved equal sensitivity but higher accuracy and specificity than PlasFlow, while the predictions made by Platon were more balanced in terms of F1 score and MCC due to a low number of false positives.

Even though the underlying MPS database follows an untargeted approach, i.e. it is not restricted to or focused on certain taxa, Platon achieved competitive results compared to the targeted tools PlaScope and PlasmidFinder in a benchmark using real sequencing data for *E. coli* isolates. In both benchmarks Platon achieved the highest sensitivity and accuracy, thus endorsing the exploitation of the natural replicon distribution biases of protein-coding genes as an eligible method for the large-scale, high-throughput, taxon-independent prediction of plasmid-borne contigs from short-read draft assemblies.

Implemented as a multithreaded, locally executable Linux command line application in Python 3, we also envision it as an appropriate fit for integration into larger analysis pipelines as well as an upfront tool for subsequent plasmid-specific analyses. For the sake of a streamlined integration and installation, all necessary third party executables are bundled with the software. All source code and documentation are freely available under a GPL3 license and hosted at GitHub (https://github.com/oschwengers/platon) and http://platon.computational.bio/. For further convenience, Platon is also available as a BioConda package (platon) and via PyPI (cb-platon). A prebuilt database is hosted at Zenodo (DOI: 10.5281/zenodo.3349652).

Future developments will include the addition of new higher-level contig characterizations as well as further enhancements of applied heuristics.

Funding information

This work was supported by the German Center of Infection Research (DZIF) [DZIF grant 8000701–3 (HZI), TI06.001, 8032808811 and 8032808820 to T. C.]; the German Network for Bioinformatics Infrastructure (de.NBI) (BMBF grant FKZ 031A533B to A. G.); and the German Research Foundation (DFG) [SFB-TR84 project A04 (TRR84/3 2018) to T. C., KFO309 Z1 (GO 2037/5-1) to A. G., SFB-TR84 project B08 (TRR84/3 2018) to T. H., SFB1021 Z02 (SFB 1021/2 2017) to T. H. and KFO309 Z1 (HA 5225/1–1) to T. H.].

Acknowledgements

The authors thank Jane Falgenhauer for her valuable ideas, testing, bug reports and fruitful feedback. Furthermore, the authors thank Christina Gerstmann for excellent technical assistance.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Data Bibliography

1. Platon was developed as a Python 3 command line application for Linux.

2. The complete source code and documentation are available on GitHub under a GPL3 license: https://github.com/oschwengers/platon and http://platon.computational.bio.

3. All database versions are hosted at Zenodo (DOI: 10.5281/zenodo.3349651).

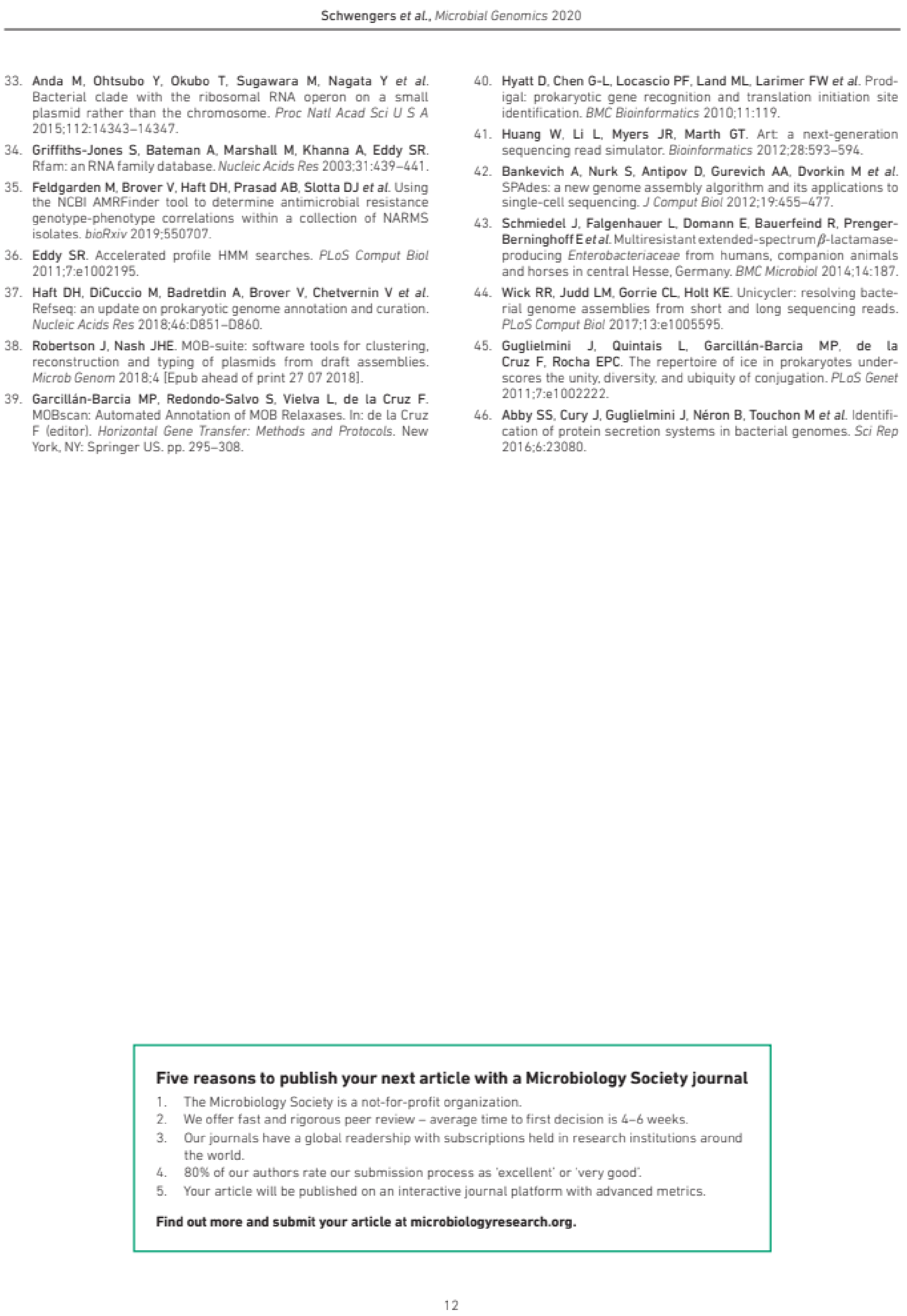
4. Platon is available via the bioconda package platon.

5. Platon is available via the PyPI package cb-platon.

6. The bacterial representative sequences for UniProt's UniRef90 protein clusters, complete bacterial genome sequences from the NCBI RefSeq database, complete plasmid sequences from the NCBI genomes plasmid section, created artificial contigs, RDS threshold metrics and raw protein replicon hit counts used to create and evaluate the marker protein sequence database are hosted at Zenodo (DOI: 10.5281/zenodo.37591697). Twenty-four *Escherichia coli* isolates sequenced with short-read (Illumina MiSeq) and long-read sequencing technologies (Oxford Nanopore Technology GridION platform) used for real data benchmarks are available under the following NCBI BioProjects: PRJNA505407 and PRJNA387731.

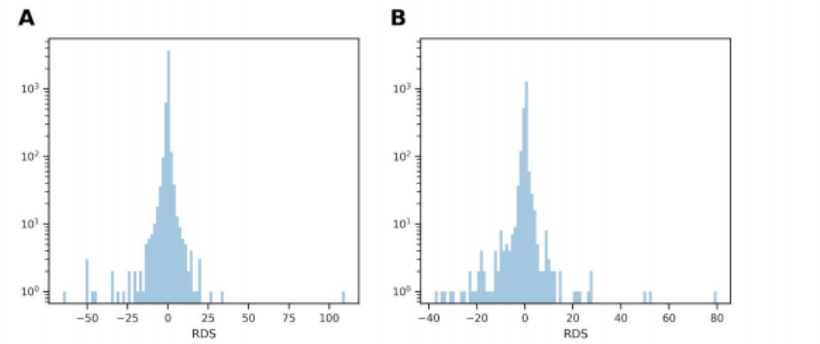
References

1. **ClarkDP, StahlDA, MartinkoJM, MadiganMT**. 2010. Brock biology of microorganisms (13th edition). Benjamin Cummings. https://www.amazon.com/Brock-Biology-Microorganisms-Michael-Madigan/dp/032164963X
2. **Tazzyman SJ, Bonhoeffer S**. Why there are no essential genes on plasmids. *Mol Biol Evol* 2015;32:3079–3088.
3. **Thomas CM, Nielsen KM**. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 2005;3:711–721.
4. **Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F**. Mobility of plasmids. *Microbiol Mol Biol Rev* 2010;74:434–452.
5. **Carattoli A**. Plasmids and the spread of resistance. *Int J Med Microbiol* 2013;303:298–304.
6. **Dierikx C, van der Goot J, Fabri T, van Essen-Zandbergen A, Smith H** *et al*. Extended-spectrum-*β*-lactamase- and AmpC-*β*-lactamase-producing *Escherichia coli* in Dutch broilers and broiler farmers. *J Antimicrob Chemother* 2013;68:60–67.
7. **Schweizer C, Bischoff P, Bender J, Kola A, Gastmeier P** *et al*. Plasmid-Mediated Transmission of KPC-2 Carbapenemase in *Enterobacteriaceae* in Critically Ill Patients. *Front Microbiol* 2019;10:276.
8. **Zheng R, Zhang Q, Guo Y, Feng Y, Liu L** *et al*. Outbreak of plasmid-mediated NDM-1-producing *Klebsiella pneumoniae* ST105 among neonatal patients in Yunnan, China. *Ann Clin Microbiol Antimicrob* 2016;15:10.
9. **Yie Y, Wei Z, Tien P**. A simplified and reliable protocol for plasmid DNA sequencing: fast miniprep and denaturation. *Nucleic Acids Res* 1993;21:361.
10. **Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ** *et al*. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front Microbiol* 2017;8:182.
11. **Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC**. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3:1–8.
12. **Cohen SN**. Transposable genetic elements and plasmid evolution. *Nature* 1976;263:731–734.
13. **Escudero JA, Loot C, Nivina A, Mazel D**. The integron: adaptation on demand. *Microbiol Spectr* 2015;3:MDNA3–0019–2014.
14. **Sohn J-I, Nam J-W**. The present and future of de novo whole-genome assembly. *Brief Bioinform* 2018;19:23–40.
15. **Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E** *et al*. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics* 2016;95:btw651.
16. **Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A** *et al*. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016;32:btw493–3387.
17. **Vielva L, de Toro M, Lanza VF, de la Cruz F**. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics* 2017;33:3796–3798.
18. **Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O** *et al*. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.
19. **Zhou F, Xu Y**. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 2010;26:2051–2052.
20. **Krawczyk PS, Lipinski L, Dziembowski A**. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;46:e35.
21. **Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J** *et al*. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom* 2018;4:1–15.
22. **Royer G, Decousser JW, Branger C, Dubois M, Médigue C** *et al*. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom* 2018;4.
23. **Roosaare M, Puustusmaa M, Möls M, Vaher M, Remm M**. PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ* 2018;6:e4588.
24. **Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C** *et al*. Big data: astronomical or Genomical? *PLoS Biol* 2015;13:e1002195.
25. **Caboche S, Even G, Loywick A, Audebert C, Hot D**. MICRA: an automatic pipeline for fast characterization of microbial genomes from high-throughput sequencing data. *Genome Biol* 2017;18:233.
26. **Quijada NM, Rodríguez-Lázaro D, Eiros JM, Hernández M**. TORMES: an automated pipeline for whole bacterial genome analysis. *Bioinformatics* 2019;35:4207–4212.
27. **Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T** *et al*. ASA3P: an automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. *PLoS Comput Biol* 2020;16:e1007134.
28. **Galata V, Fehlmann T, Backes C, Keller A**. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res* 2019;47:D195–D202.
29. **UniProt Consortium**. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;47:D506–D515.
30. **Buchfink B, Xie C, Huson DH**. Fast and sensitive protein alignment using diamond. *Nat Methods* 2015;12:59–60.
31. **Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL** *et al*. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* 2018;14:e1005944.
32. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J** *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.

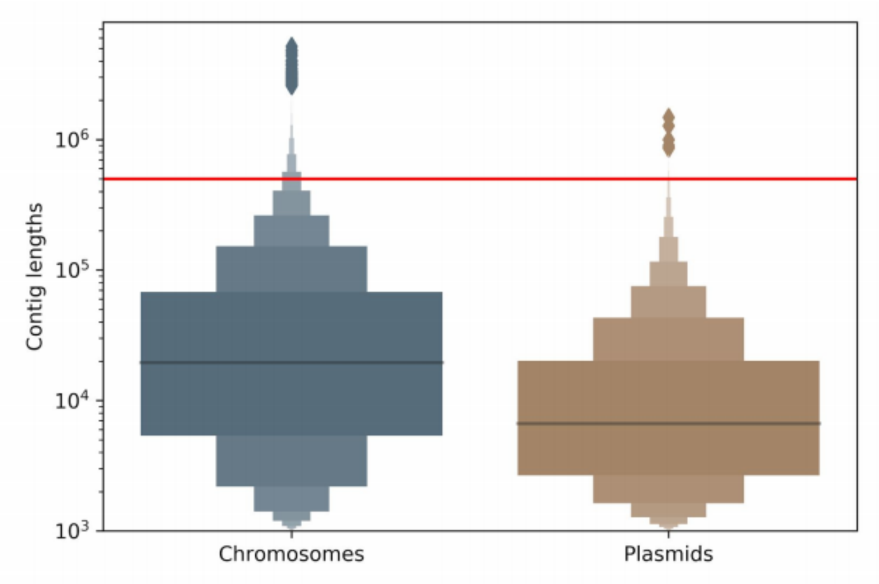


33. **Anda M, Ohtsubo Y, Okubo T, Sugawara M, Nagata Y** *et al.* Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. *Proc Natl Acad Sci U S A* 2015;112:14343–14347.
34. **Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR.** Rfam: an RNA family database. *Nucleic Acids Res* 2003;31:439–441.
35. **Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ** *et al.* Using the NCBI AMRFinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of NARMS isolates. *bioRxiv* 2019;550707.
36. **Eddy SR.** Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.
37. **Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V** *et al.* Refseq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res* 2018;46:D851–D860.
38. **Robertson J, Nash JHE.** MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4 [Epub ahead of print 27 07 2018].
39. **Garcillán-Barcia MP, Redondo-Salvo S, Vielva L, de la Cruz F.** MOBscan: Automated Annotation of MOB Relaxases. In: de la Cruz F (editor). *Horizontal Gene Transfer: Methods and Protocols*. New York, NY: Springer US. pp. 295–308.
40. **Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW** *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
41. **Huang W, Li L, Myers JR, Marth GT.** Art: a next-generation sequencing read simulator. *Bioinformatics* 2012;28:593–594.
42. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M** *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
43. **Schmiedel J, Falgenhauer L, Domann E, Bauerfeind R, Prenger-Berninghoff E** *et al.* Multiresistant extended-spectrum β-lactamase-producing *Enterobacteriaceae* from humans, companion animals and horses in central Hesse, Germany. *BMC Microbiol* 2014;14:187.
44. **Wick RR, Judd LM, Gorrie CL, Holt KE.** Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
45. **Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC.** The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet* 2011;7:e1002222.
46. **Abby SS, Cury J, Guglielmini J, Néron B, Touchon M** *et al.* Identification of protein secretion systems in bacterial genomes. *Sci Rep* 2016;6:23080.

7.4 Supplementary Information – Platon

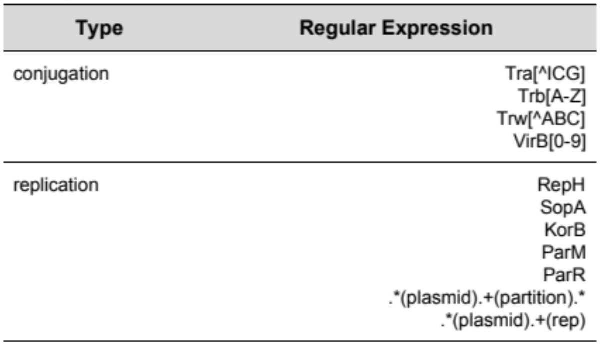


Supplementary Figure S1. Histogram of RDS values for (A) relaxase and (B) type 4-coupling proteins.



Supplementary Figure S2. Length distribution of chromosome and plasmid contigs resulted from simulated short-read assemblies. Outliers are shown as diamonds; horizontal red line: implemented contig length heuristic threshold (n=500,000 bp) as applied in the platon workflow.

Supplementary Table S1. Regular expressions used for PCLA cluster extraction for subsequent HMM creation



Type	Regular Expression
conjugation	Tra[^ICG] Trb[A-Z] Trw[^ABC] VirB[0-9]
replication	RepH SopA KorB ParM ParR .*(plasmid).+(partition).* .*(plasmid).+(rep)

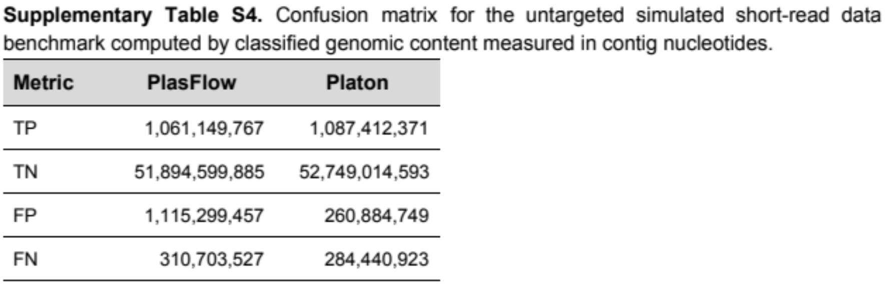
Supplementary Table S2. Isolated and sequenced *Escherichia coli* genomes used in the real data benchmark.

Isolate	SRA Accession ID	Host	Assembled Contigs from Short Reads >= 1 kbp	Chromosome Closed in Hybrid Assemblies	Closed Plasmids in Hybrid Assemblies
H69	SRX5007771 SRX5007759	Homo sapiens	69	yes	2
H100	SRX5007774 SRX5007760	Homo sapiens	44	yes	3
H105	SRX5002893 SRX5002892	Homo sapiens	59	yes	4
H108	SRX5007773 SRX5007761	Homo sapiens	69	yes	3
H113	SRX5007776 SRX5007762	Homo sapiens	35	yes	2
H136	SRX5007775 SRX5007763	Homo sapiens	77	yes	5
H157	SRX5007770 SRX5007756	Homo sapiens	87	yes	4
H162	SRX5007769 SRX5007757	Homo sapiens	56	yes	1
H176	SRX5007772 SRX5007758	Homo sapiens	92	yes	2
V1	SRX5007768 SRX5007764	Canis lupus	61	yes	6
V8	SRX5007790 SRX5007782	Equus caballus	46	yes	2
V9	SRX5007786 SRX5007784	Equus caballus	118	yes	5
V41	SRX5007794 SRX5007777	Canis lupus	82	yes	3
V64	SRX5007789 SRX5007780	Canis lupus	52	yes	2
V71	SRX5007788 SRX5007766	Canis lupus	90	yes	7
V73	SRX5007791 SRX5007783	Equus caballus	65	yes	2
V79	SRX6897800	Equus caballus	55	no	5

	SRX6897801				
V80	SRX5007787 SRX5007785	Equus caballus	68	yes	0
V173	SRX5007767 SRX5007765	Equus caballus	75	yes	1
V177	SRX5007793 SRX5007781	Canis lupus	66	yes	3
V195	SRX5007792 SRX5007779	Canis lupus	51	yes	4
V292	SRX5007795 SRX5007778	Canis lupus	99	yes	3
V215	SRX6897802 SRX6897803	Equus caballus		no	6
V232	SRX6897804 SRX6897805	Canis lupus		no	1

Supplementary Table S3. Number of classified contigs for each RDS and length threshold and characterization heuristic implemented in the Platon workflow for both simulated and real benchmarks.

RDS threshold / heuristic	Simulated data	Real data
Length < 1kb	0	0
Length >= 500 kb	15,869	22
RDS SNT	443,159	985
RDS SPT	42,669	241
RDS CT	50,525	212
Circularity	53,611	266
Incompatibility group	5,749	82
Replication gene	6,772	48
Mobilization gene	287	0
OriT	1,614	40



Supplementary Table S4. Confusion matrix for the untargeted simulated short-read data benchmark computed by classified genomic content measured in contig nucleotides.

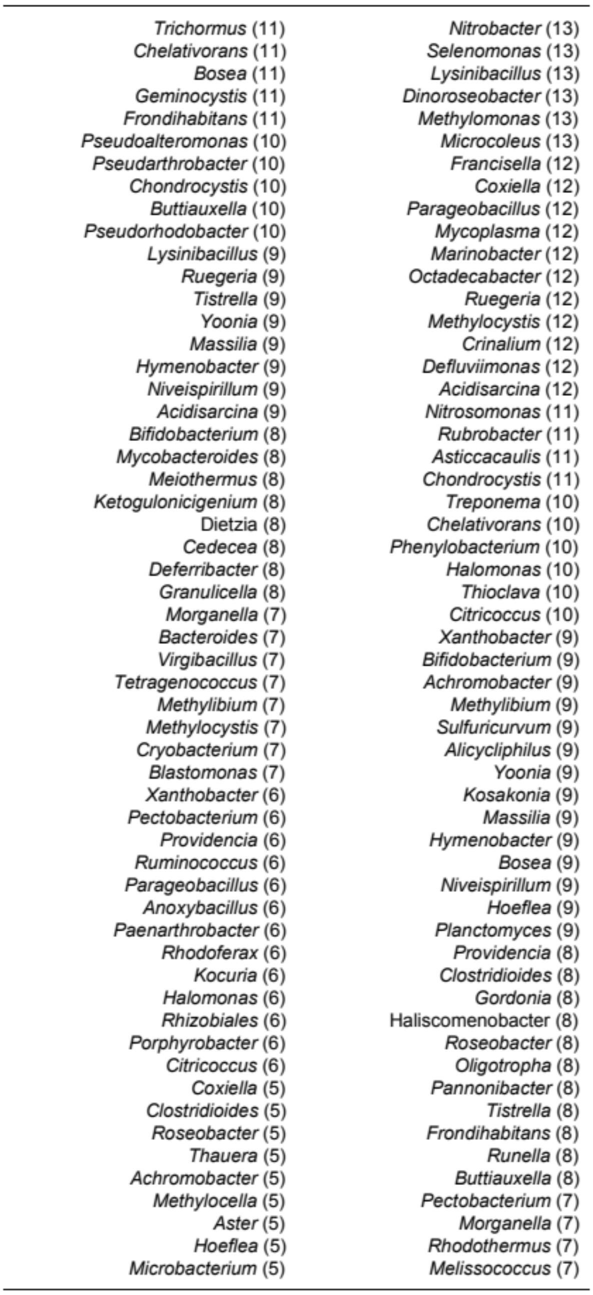
Metric	PlasFlow	Platon
TP	1,061,149,767	1,087,412,371
TN	51,894,599,885	52,749,014,593
FP	1,115,299,457	260,884,749
FN	310,703,527	284,440,923

Supplementary Table S5. Taxa of bacterial genomes for which true plasmid contigs have been correctly identified by each tool in the simulated short-read benchmark binned to the *genus* taxon. Aggregated counts for each *genus* are provided in parenthesis.

PlasFlow	Platon
Escherichia (7585)	*Klebsiella* (7651)
Klebsiella (7577)	*Escherichia* (7099)
Enterococcus (2309)	*Enterococcus* (3382)
Bacillus (2150)	*Bacillus* (2109)
Salmonella (2002)	*Salmonella* (1835)
Sinorhizobium (1713)	*Sinorhizobium* (1706)
Rhizobium (1707)	*Rhizobium* (1684)
Ralstonia (1633)	*Ralstonia* (1659)
Lactobacillus (1139)	*Lactobacillus* (1596)
Shigella (1124)	*Shigella* (1109)
Enterobacter (898)	*Enterobacter* (881)
Xanthomonas (653)	*Acinetobacter* (792)
Acinetobacter (647)	*Piscirickettsia* (750)
Staphylococcus (645)	*Acetobacter* (728)
Acetobacter (591)	*Xanthomonas* (641)
Citrobacter (570)	*Staphylococcus* (638)
Pseudomonas (534)	*Citrobacter* (571)
Azospirillum (504)	*Pseudomonas* (539)
Piscirickettsia (489)	*Borreliella* (503)
Yersinia (456)	*Yersinia* (464)
Burkholderia (433)	*Phaeobacter* (429)
Borreliella (336)	*Azospirillum* (418)
Vibrio (326)	*Borrelia* (380)
Sphingobium (319)	*Burkholderia* (363)
Rhodococcus (296)	*Vibrio* (339)
Borrelia (292)	*Sphingobium* (310)
Lactococcus (285)	*Lactococcus* (308)
Phaeobacter (283)	*Rhodococcus* (258)
Microvirga (260)	*Streptomyces* (258)
Agrobacterium (259)	*Paracoccus* (249)
Paracoccus (234)	*Agrobacterium* (242)
Paraburkholderia (232)	*Deinococcus* (216)
Streptomyces (208)	*Aeromonas* (200)
Aeromonas (183)	*Microvirga* (165)
Deinococcus (173)	*Clostridium* (163)
Nostoc (162)	*Nostoc* (153)
Cupriavidus (161)	*Mycobacterium* (145)
Sphingomonas (157)	*Cupriavidus* (145)
Rhodobacter (151)	*Pantoea* (139)
Mycobacterium (151)	*Pediococcus* (135)
Pantoea (143)	*Pseudonocardia* (128)
Methylobacterium (142)	*Campylobacter* (121)
Novosphingobium (128)	*Novosphingobium* (120)
Komagataeibacter (123)	*Thermus* (118)
Bradyrhizobium (114)	*Komagataeibacter* (118)
Leclercia (112)	*Sphingomonas* (114)
Shewanella (103)	*Rhodobacter* (113)
Raoultella (101)	*Paraburkholderia* (108)

Pediococcus (99)	*Raoultella* (106)
Ensifer (96)	*Moraxella* (102)
Pandoraea (91)	*Shewanella* (102)
Clostridium (89)	*Leuconostoc* (96)
Pseudonocardia (85)	*Ensifer* (96)
Thermus (82)	*Sulfitobacter* (95)
Sulfitobacter (80)	*Leclercia* (92)
Acaryochloris (75)	*Methylobacterium* (91)
Ochrobactrum (69)	*Bradyrhizobium* (91)
Campylobacter (68)	Listeria (82)
Acidiphilium (67)	*Arsenophonus* (79)
Serratia (66)	*Legionella* (77)
Mesorhizobium (62)	*Gloeothece* (77)
Gloeothece (62)	*Acaryochloris* (76)
Roseomonas (61)	*Pandoraea* (73)
Photobacterium (60)	*Candidatus* (71)
Arsenophonus (59)	*Sphingopyxis* (67)
Methylorubrum (57)	*Synechococcus* (65)
Sphingopyxis (57)	*Paenibacillus* (62)
Kozakia (55)	*Ochrobactrum* (60)
Arthrobacter (54)	*Arthrobacter* (60)
Leptolyngbya (51)	*Chlamydia* (59)
Leuconostoc (50)	*Photobacterium* (59)
Enterobacteriaceae (49)	*Serratia* (58)
Shinella (48)	*Acidiphilium* (57)
Phytobacter (48)	*Synechocystis* (57)
Listeria (47)	*Leptolyngbya* (56)
Proteus (44)	*Azotobacter* (51)
Cronobacter (43)	*Kozakia* (51)
Candidatus (40)	*Cronobacter* (50)
Gluconobacter (39)	*Zymomonas* (48)
Mycoplasma (39)	*Enterobacteriaceae* (47)
Synechococcus (38)	*Aminobacter* (46)
Moraxella (37)	*Phytobacter* (46)
Synechocystis (36)	*Proteus* (43)
Haematobacter (36)	*Methylorubrum* (42)
Helicobacter (35)	*Clavibacter* (41)
Aminobacter (33)	*Shinella* (41)
Paenibacillus (32)	*Mesorhizobium* (40)
Acidovorax (32)	*Acidovorax* (37)
Corynebacterium (31)	*Roseomonas* (37)
Mycolicibacterium (31)	*Geobacillus* (36)
Streptococcus (30)	*Corynebacterium* (36)
Neorhizobium (29)	*Gluconobacter* (35)
Carnobacterium (29)	*Neisseria* (34)
Croceicoccus (29)	*Leptospira* (33)
Azotobacter (28)	*Weissella* (31)
Geobacillus (28)	*Acidithiobacillus* (31)
Erwinia (26)	*Rickettsia* (31)
Celeribacter (25)	*Helicobacter* (30)
Antarctobacter (24)	*Haematobacter* (30)
Legionella (23)	*Psychrobacter* (29)
Aliivibrio (23)	*Meiothermus* (29)
Oscillatoria (23)	*Leisingera* (29)
Chlamydia (22)	*Celeribacter* (29)
Xylella (22)	*Planococcus* (28)

Stanieria (22)	*Calothrix* (28)
Aureimonas (22)	*Stanieria* (28)
Confluentimicrobium (22)	*Polaromonas* (28)
Phyllobacterium (22)	*Erwinia* (27)
Zymomonas (21)	*Carnobacterium* (27)
Xenorhabdus (21)	*Oscillatoria* (27)
Nitrobacter (21)	*Streptococcus* (26)
Rahnella (20)	*Pseudanabaena* (26)
Rhodovulum (20)	*Xylella* (25)
Octadecabacter (20)	*Antarctobacter* (25)
Aromatoleum (20)	*Yangia* (25)
Rhizorhabdus (20)	*Indioceanicola* (25)
Neisseria (19)	*Rhodovulum* (24)
Gordonia (19)	*Croceicoccus* (24)
Yangia (19)	*Neorhizobium* (23)
Epibacterium (19)	*Mycolicibacterium* (23)
Buchnera (18)	*Martelella* (23)
Martelella (18)	*Cyanothece* (23)
Indioceanicola (18)	*Bacteroides* (22)
Rippkaea (18)	*Ruminococcus* (22)
Leptospira (17)	*Rippkaea* (22)
Weissella (17)	*Buchnera* (21)
Nocardia (17)	*Aliivibrio* (21)
Psychrobacter (17)	*Anabaena* (21)
Thioclava (17)	*Granulicella* (20)
Edwardsiella (16)	*Confluentimicrobium* (20)
Anabaena (16)	*Pseudoalteromonas* (19)
Cyanothece (16)	*Marinovum* (19)
Alteromonas (15)	*Deferribacter* (19)
Rickettsia (15)	*Phyllobacterium* (19)
Polaromonas (15)	*Methylosinus* (18)
Sagittula (15)	*Ilyobacter* (18)
Methylosinus (14)	*Aromatoleum* (18)
Planococcus (14)	*Salipiger* (18)
Marinobacter (14)	*Aureimonas* (18)
Calothrix (14)	*Geminocystis* (18)
Leisingera (14)	*Peptoclostridium* (17)
Dinoroseobacter (14)	*Ketogulonicigenium* (17)
Salipiger (14)	*Epibacterium* (17)
Pelagibaca (14)	*Pelagibaca* (17)
Defluviimonas (14)	*Xenorhabdus* (16)
Spiroplasma (13)	*Desulfovibrio* (16)
Clavibacter (13)	*Rhizorhabdus* (16)
Microcoleus (13)	*Sedimentitalea* (16)
Sedimentitalea (13)	*Edwardsiella* (15)
Metakosakonia (13)	*Rahnella* (15)
Marinovum (12)	*Pseudarthrobacter* (15)
Pseudanabaena (12)	*Sagittula* (15)
Alicycliphilus (12)	*Pseudorhodobacter* (15)
Gemmobacter (12)	*Alteromonas* (14)
Francisella (11)	*Nocardia* (14)
Acidithiobacillus (11)	*Salinibacter* (14)
Oligotropha (11)	*Gemmatirosa* (14)
Macrococcus (11)	*Gemmobacter* (14)
Kosakonia (11)	*Metakosakonia* (14)
Crinalium (11)	*Fusobacterium* (13)



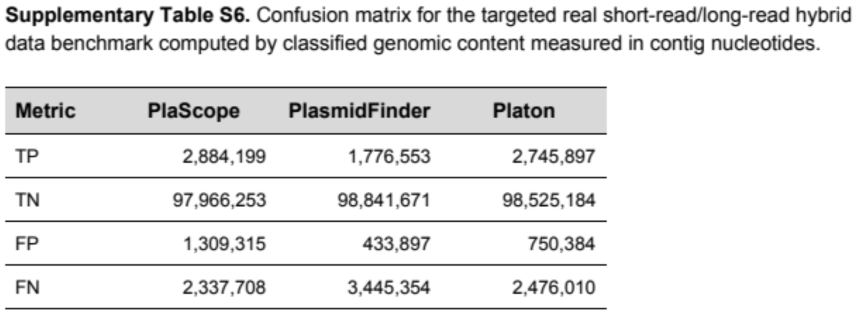
Trichormus (11)	*Nitrobacter* (13)
Chelativorans (11)	*Selenomonas* (13)
Bosea (11)	*Lysinibacillus* (13)
Geminocystis (11)	*Dinoroseobacter* (13)
Frondihabitans (11)	*Methylomonas* (13)
Pseudoalteromonas (10)	*Microcoleus* (13)
Pseudarthrobacter (10)	*Francisella* (12)
Chondrocystis (10)	*Coxiella* (12)
Buttiauxella (10)	*Parageobacillus* (12)
Pseudorhodobacter (10)	*Mycoplasma* (12)
Lysinibacillus (9)	*Marinobacter* (12)
Ruegeria (9)	*Octadecabacter* (12)
Tistrella (9)	*Ruegeria* (12)
Yoonia (9)	*Methylocystis* (12)
Massilia (9)	*Crinalium* (12)
Hymenobacter (9)	*Defluviimonas* (12)
Niveispirillum (9)	*Acidisarcina* (12)
Acidisarcina (9)	*Nitrosomonas* (11)
Bifidobacterium (8)	*Rubrobacter* (11)
Mycobacteroides (8)	*Asticcacaulis* (11)
Meiothermus (8)	*Chondrocystis* (11)
Ketogulonicigenium (8)	*Treponema* (10)
Dietzia (8)	*Chelativorans* (10)
Cedecea (8)	*Phenylobacterium* (10)
Deferribacter (8)	*Halomonas* (10)
Granulicella (8)	*Thioclava* (10)
Morganella (7)	*Citricoccus* (10)
Bacteroides (7)	*Xanthobacter* (9)
Virgibacillus (7)	*Bifidobacterium* (9)
Tetragenococcus (7)	*Achromobacter* (9)
Methylibium (7)	*Methylibium* (9)
Methylocystis (7)	*Sulfuricurvum* (9)
Cryobacterium (7)	*Alicycliphilus* (9)
Blastomonas (7)	*Yoonia* (9)
Xanthobacter (6)	*Kosakonia* (9)
Pectobacterium (6)	*Massilia* (9)
Providencia (6)	*Hymenobacter* (9)
Ruminococcus (6)	*Bosea* (9)
Parageobacillus (6)	*Niveispirillum* (9)
Anoxybacillus (6)	*Hoeflea* (9)
Paenarthrobacter (6)	*Planctomyces* (9)
Rhodoferax (6)	*Providencia* (8)
Kocuria (6)	*Clostridioides* (8)
Halomonas (6)	*Gordonia* (8)
Rhizobiales (6)	Haliscomenobacter (8)
Porphyrobacter (6)	*Roseobacter* (8)
Citricoccus (6)	*Oligotropha* (8)
Coxiella (5)	*Pannonibacter* (8)
Clostridioides (5)	*Tistrella* (8)
Roseobacter (5)	*Frondihabitans* (8)
Thauera (5)	*Runella* (8)
Achromobacter (5)	*Buttiauxella* (8)
Methylocella (5)	*Pectobacterium* (7)
Aster (5)	*Morganella* (7)
Hoeflea (5)	*Rhodothermus* (7)
Microbacterium (5)	*Melissococcus* (7)

Rhodobacteraceae (5)	*Myroides* (7)
Acidibrevibacterium (5)	*Chryseobacterium* (7)
Tabrizicola (5)	*Dietzia* (7)
Crocosphaera (5)	*Trichormus* (7)
Beijerinckia (4)	*Kocuria* (7)
Plesiomonas (4)	*Acidibrevibacterium* (7)
Actinobacillus (4)	*Virgibacillus* (6)
Pasteurella (4)	*Desulfobacterium* (6)
Nitrosomonas (4)	*Arcobacter* (6)
Thiomonas (4)	*Lawsonia* (6)
Selenomonas (4)	*Paenarthrobacter* (6)
Allochromatium (4)	*Tetragenococcus* (6)
Microcystis (4)	*Sodalis* (6)
Brevibacillus (4)	*Caulobacter* (6)
Kitasatospora (4)	*Macrococcus* (6)
Tsukamurella (4)	*Rhodoferax* (6)
Lawsonia (4)	*Simkania* (6)
Gluconacetobacter (4)	*Amycolatopsis* (6)
Glutamicibacter (4)	*Cedecea* (6)
Nodularia (4)	*Thalassospira* (6)
Myroides (4)	*Methylocella* (6)
Asticcacaulis (4)	*Singulisphaera* (6)
Desulfovibrio (4)	*Aquabacterium* (6)
Labrenzia (4)	*Rhodobacteraceae* (6)
Phenylobacterium (4)	*Clostridiaceae* (6)
Exiguobacterium (4)	*Lelliottia* (6)
Oscillibacter (4)	*Tabrizicola* (6)
Vagococcus (4)	*Plesiomonas* (5)
Neokomagataea (4)	*Nitrosococcus* (5)
Plautia (4)	*Nitrosospira* (5)
Swingsia (4)	*Pelobacter* (5)
Hartmannibacter (4)	*Mycobacteroides* (5)
Clostridiaceae (4)	*Butyrivibrio* (5)
Simplicispira (4)	*Chroococcidiopsis* (5)
Bordetella (3)	*Labrenzia* (5)
Hafnia (3)	*Roseovarius* (5)
Fusobacterium (3)	*Tateyamaria* (5)
Rhodospirillum (3)	*Exiguobacterium* (5)
Halobacillus (3)	*Oscillibacter* (5)
Desulfobacterium (3)	*Flammeovirga* (5)
Bartonella (3)	*Thermaerobacter* (5)
Sodalis (3)	*Porphyrobacter* (5)
Pannonibacter (3)	*Brachyspira* (4)
Salinibacter (3)	*Hydrogenophilus* (4)
Roseovarius (3)	*Thiomonas* (4)
Thalassospira (3)	*Rhodospirillum* (4)
Alicyclobacillus (3)	*Halobacillus* (4)
Chelatococcus (3)	*Prevotella* (4)
Salimicrobium (3)	*Geobacter* (4)
Haematospirillum (3)	*Glutamicibacter* (4)
Erythrobacter (3)	*Thauera* (4)
Nostocales (3)	*Curtobacterium* (4)
Glaesserella (3)	*Thioflavicoccus* (4)
Lelliottia (3)	*Bartonella* (4)
Thiomicrorhabdus (3)	*Blattabacterium* (4)
Hydrocarboniclastica (3)	*Geoalkalibacter* (4)

Vitreoscilla (2)	*Neokomagataea* (4)
Hydrogenophilus (2)	*Cryobacterium* (4)
Sebaldella (2)	*Azoarcus* (4)
Finegoldia (2)	*Gloeocapsa* (4)
Cutibacterium (2)	*Haematospirillum* (4)
Haliscomenobacter (2)	*Blastomonas* (4)
Brochothrix (2)	*Cnuibacter* (4)
Pelobacter (2)	*Microbacterium* (4)
Caldicellulosiruptor (2)	*Actinobacillus* (3)
Desulfohalobium (2)	*Allochromatium* (3)
Prevotella (2)	*Microcystis* (3)
Chroococcidiopsis (2)	*Brevibacillus* (3)
Thioflavicoccus (2)	*Tsukamurella* (3)
Moritella (2)	*Spiroplasma* (3)
Simkania (2)	*Caldicellulosiruptor* (3)
endosymbiont (2)	*Desulfohalobium* (3)
Caulobacter (2)	*Nodularia* (3)
Rivularia (2)	*Aster* (3)
Cyanobacterium (2)	*Thermovirga* (3)
Singulisphaera (2)	*Thermobacillus* (3)
Hoyosella (2)	*Cyanobacterium* (3)
Azoarcus (2)	*Vagococcus* (3)
Polymorphum (2)	*Salimicrobium* (3)
Dickeya (2)	*Neochlamydia* (3)
Gloeocapsa (2)	*Simplicispira* (3)
Aquabacterium (2)	*Chromobacterium* (3)
Halocynthiibacter (2)	*Thiomicrorhabdus* (3)
Euzebya (2)	*Silvanigrellales* (3)
Cnuibacter (2)	*Crocosphaera* (3)
Planctomyces (2)	*Vitreoscilla* (2)
Nitratireductor (2)	*Bordetella* (2)
Brachybacterium (2)	*Beijerinckia* (2)
Gammaproteobacteria (2)	*Hafnia* (2)
Sterolibacteriaceae (2)	*Pasteurella* (2)
Thermaerobacter (2)	*Sebaldella* (2)
Comamonas (1)	*Finegoldia* (2)
Alcaligenes (1)	*Dermacoccus* (2)
Histophilus (1)	*Streptosporangium* (2)
Gallibacterium (1)	*Gluconacetobacter* (2)
Marivirga (1)	*Streptobacillus* (2)
Rhodopseudomonas (1)	*Sinomonas* (2)
Prosthecochloris (1)	*Tatumella* (2)
Nitrosospira (1)	*Pseudodesulfovibrio* (2)
Dermacoccus (1)	*Desulfocapsa* (2)
Brevibacterium (1)	*endosymbiont* (2)
Peptoclostridium (1)	*Cardinium* (2)
Acidipropionibacterium (1)	*Methylovorus* (2)
Desulfurella (1)	*Jannaschia* (2)
Zymobacter (1)	*Anoxybacillus* (2)
Sinomonas (1)	*Advenella* (2)
Rubrobacter (1)	*Rivularia* (2)
Hydrogenophaga (1)	*Natranaerobius* (2)
Wigglesworthia (1)	*Thioalkalivibrio* (2)
Tatumella (1)	*Maritalea* (2)
Waddlia (1)	*Calditerrivibrio* (2)
Mannheimia (1)	*Rufibacter* (2)

Solibacillus (1)	*Opitutaceae* (2)
Brachyspira (1)	*Desulfosporosinus* (2)
Desulfotalea (1)	*Plautia* (2)
Halobacteriovorax (1)	*Halioglobus* (2)
Kineococcus (1)	*Polymorphum* (2)
Rummeliibacillus (1)	*Dickeya* (2)
Cardinium (1)	*Altererythrobacter* (2)
Methylovorus (1)	*Capnocytophaga* (2)
Jannaschia (1)	*Paludisphaera* (2)
Photorhabdus (1)	*Cetia* (2)
Tateyamaria (1)	*Hartmannibacter* (2)
Advenella (1)	*Rickettsiales* (2)
Geobacter (1)	*Erythrobacter* (2)
Verminephrobacter (1)	*Nostocales* (2)
Natranaerobius (1)	*Thalassococcus* (2)
Thioalkalivibrio (1)	*Glaesserella* (2)
Tessaracoccus (1)	*Gammaproteobacteria* (2)
Nitrosococcus (1)	*Catenovulum* (2)
Prauserella (1)	*Humibacter* (2)
Allofrancisella (1)	*Planctopirus* (1)
Frankia (1)	*Isosphaera* (1)
Jeotgalibaca (1)	*Comamonas* (1)
Methylophaga (1)	*Alcaligenes* (1)
Arcobacter (1)	*Histophilus* (1)
Pusillimonas (1)	*Gallibacterium* (1)
Moorea (1)	*Herbaspirillum* (1)
Catharanthus (1)	*Marivirga* (1)
Mucilaginibacter (1)	*Saprospira* (1)
Cycloclasticus (1)	*Prosthecochloris* (1)
Paludisphaera (1)	*Gottschalkia* (1)
Geosporobacter (1)	*Kitasatospora* (1)
Sedimenticola (1)	*Hirschia* (1)
Aquitalea (1)	*Brochothrix* (1)
Psychromicrobium (1)	*Turneriella* (1)
Spongiibacter (1)	*Desulfurella* (1)
Magnetospirillum (1)	*Zymobacter* (1)
Agarilytica (1)	*Eubacterium* (1)
Fischerella (1)	*Rothia* (1)
Paraphotobacterium (1)	*Hydrogenophaga* (1)
Sphingosinicella (1)	*Wigglesworthia* (1)
Amycolatopsis (1)	*Flavobacterium* (1)
Tenericutes (1)	*Waddlia* (1)
Marivivens (1)	*Mannheimia* (1)
Sporosarcina (1)	*Moritella* (1)
Sulfuriferula (1)	*Desulfotalea* (1)
Thalassococcus (1)	*Halobacteriovorax* (1)
Ahniella (1)	*Kineococcus* (1)
Mycetocola (1)	*Marinitoga* (1)
Butyricimonas (1)	*Carboxydocella* (1)
Miniimonas (1)	*Xylanimonas* (1)
Catenovulum (1)	*Thermovibrio* (1)
Runella (1)	*Methylomicrobium* (1)
Humibacter (1)	*Collimonas* (1)
Flammeovirga (1)	*Photorhabdus* (1)
Xylanibacterium (1)	*Persephonella* (1)
Xanthomonadaceae (1)	*Pontibacter* (1)

Rhodopseudomonas (1)	*Verminephrobacter* (1)
Tatumella (1)	*Alicyclobacillus* (1)
Nitrosospira (1)	*Oceanimonas* (1)
Fischerella (1)	*Prauserella* (1)
Mannheimia (1)	*Pelagibacterium* (1)
Photorhabdus (1)	*Phycisphaera* (1)
Spongiibacter (1)	*Allofrancisella* (1)
Jannaschia (1)	*Hoyosella* (1)
Methylovorus (1)	*Sulfuricella* (1)
Thalassococcus (1)	*Frankia* (1)
Alcaligenes (1)	*Jeotgalibaca* (1)
Catenovulum (1)	*Verrucosispora* (1)
Zymobacter (1)	*Pusillimonas* (1)
Tenericutes (1)	*Thiolapillus* (1)
Xanthomonadaceae (1)	*Elizabethkingia* (1)
Mucilaginibacter (1)	*Mesotoga* (1)
Jeotgalibaca (1)	*Catharanthus* (1)
Mycetocola (1)	*Magnetospira* (1)
Psychromicrobium (1)	*Swingsia* (1)
Hydrogenophaga (1)	*Mucilaginibacter* (1)
	Cycloclasticus (1)
	Serpentinomonas (1)
	Erysipelothrix (1)
	Sedimenticola (1)
	Halocynthiibacter (1)
	Aquitalea (1)
	Psychromicrobium (1)
	Spongiibacter (1)
	Mitsuaria (1)
	Magnetospirillum (1)
	Chelatococcus (1)
	Agarilytica (1)
	Fischerella (1)
	Paraphotobacterium (1)
	Nitratireductor (1)
	Sphingosinicella (1)
	Acetobacteraceae (1)
	Marivivens (1)
	Sporosarcina (1)
	Sulfuriferula (1)
	Brachybacterium (1)
	Ahniella (1)
	Sphingorhabdus (1)
	Butyricimonas (1)
	Sterolibacteriaceae (1)
	Hydrocarboniclastica (1)
	Oenococcus (1)
	Thermoactinomycetaceae (1)
	Streptomonospora (1)
	Rhizobiales (1)



Supplementary Table S6. Confusion matrix for the targeted real short-read/long-read hybrid data benchmark computed by classified genomic content measured in contig nucleotides.

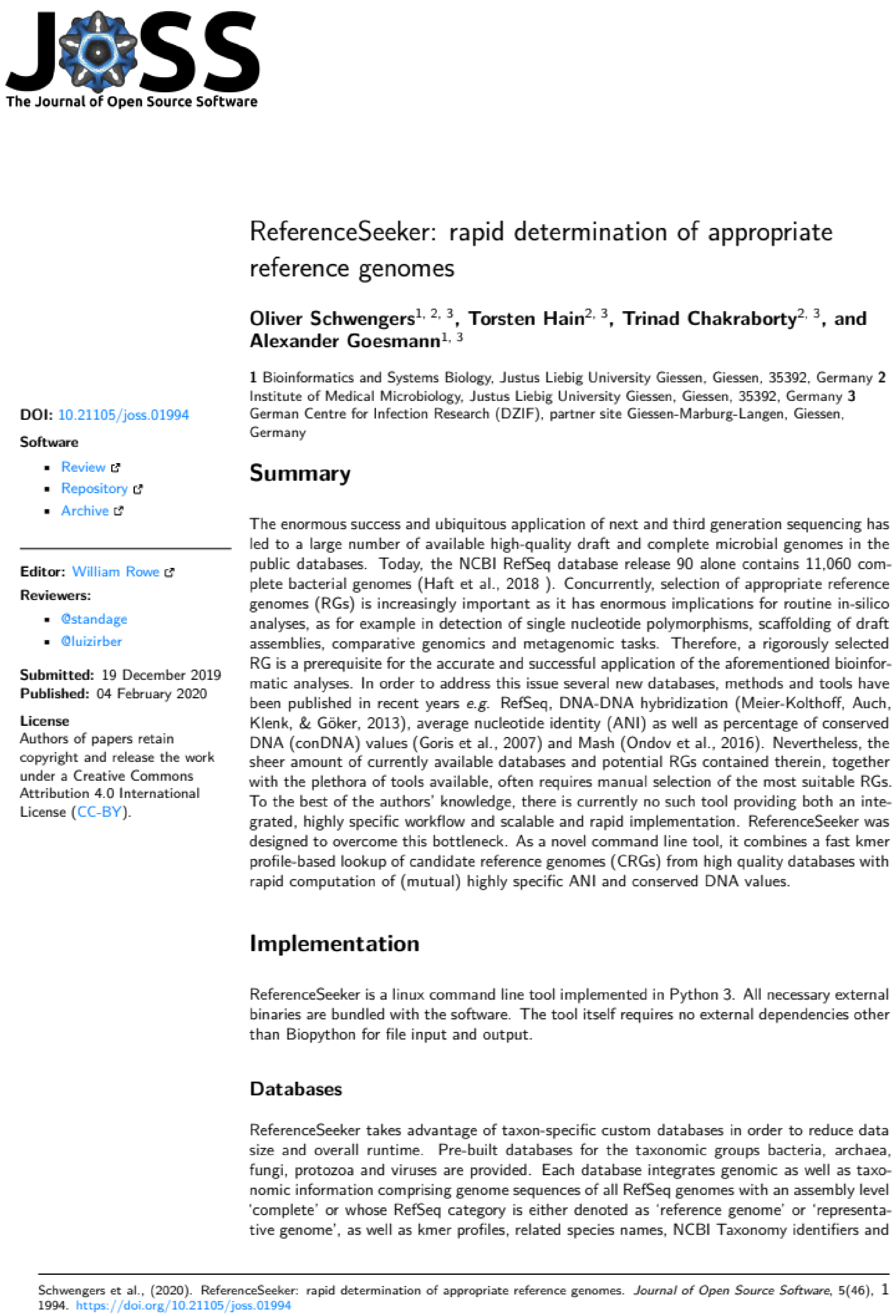
Metric	PlaScope	PlasmidFinder	Platon
TP	2,884,199	1,776,553	2,745,897
TN	97,966,253	98,841,671	98,525,184
FP	1,309,315	433,897	750,384
FN	2,337,708	3,445,354	2,476,010

7.5 ReferenceSeeker: rapid determination of appropriate reference genomes.

Oliver Schwengers[1,2,3 *], Torsten Hain[2,3], Trinad Chakraborty[2,3],
Alexander Goesmann[1,3]

Affiliations	[1] Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany [2] Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, Germany [3] German Center for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Giessen, Germany * Corresponding author

Journal	Journal of Open Source Software, Open Journals
Date	2020
Volume	5(46)
DOI	10.21105/joss.01994



ReferenceSeeker: rapid determination of appropriate reference genomes

Oliver Schwengers[1, 2, 3], **Torsten Hain**[2, 3], **Trinad Chakraborty**[2, 3], and **Alexander Goesmann**[1, 3]

1 Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, 35392, Germany **2** Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, 35392, Germany **3** German Centre for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Giessen, Germany

DOI: 10.21105/joss.01994

Software

- Review
- Repository
- Archive

Editor: William Rowe

Reviewers:

- @standage
- @luizirber

Submitted: 19 December 2019
Published: 04 February 2020

Summary

The enormous success and ubiquitous application of next and third generation sequencing has led to a large number of available high-quality draft and complete microbial genomes in the public databases. Today, the NCBI RefSeq database release 90 alone contains 11,060 complete bacterial genomes (Haft et al., 2018). Concurrently, selection of appropriate reference genomes (RGs) is increasingly important as it has enormous implications for routine in-silico analyses, as for example in detection of single nucleotide polymorphisms, scaffolding of draft assemblies, comparative genomics and metagenomic tasks. Therefore, a rigorously selected RG is a prerequisite for the accurate and successful application of the aforementioned bioinformatic analyses. In order to address this issue several new databases, methods and tools have been published in recent years *e.g.* RefSeq, DNA-DNA hybridization (Meier-Kolthoff, Auch, Klenk, & Göker, 2013), average nucleotide identity (ANI) as well as percentage of conserved DNA (conDNA) values (Goris et al., 2007) and Mash (Ondov et al., 2016). Nevertheless, the sheer amount of currently available databases and potential RGs contained therein, together with the plethora of tools available, often requires manual selection of the most suitable RGs. To the best of the authors' knowledge, there is currently no such tool providing both an integrated, highly specific workflow and scalable and rapid implementation. ReferenceSeeker was designed to overcome this bottleneck. As a novel command line tool, it combines a fast kmer profile-based lookup of candidate reference genomes (CRGs) from high quality databases with rapid computation of (mutual) highly specific ANI and conserved DNA values.

Implementation

ReferenceSeeker is a linux command line tool implemented in Python 3. All necessary external binaries are bundled with the software. The tool itself requires no external dependencies other than Biopython for file input and output.

Databases

ReferenceSeeker takes advantage of taxon-specific custom databases in order to reduce data size and overall runtime. Pre-built databases for the taxonomic groups bacteria, archaea, fungi, protozoa and viruses are provided. Each database integrates genomic as well as taxonomic information comprising genome sequences of all RefSeq genomes with an assembly level 'complete' or whose RefSeq category is either denoted as 'reference genome' or 'representative genome', as well as kmer profiles, related species names, NCBI Taxonomy identifiers and

RefSeq assembly identifiers. For convenient and fully automatic updates, we provide locally executable scripts implemented in bash and Nextflow (Di Tommaso et al., 2017). Non-public genomes can be imported into existing or newly created databases by an auxiliary command line interface.

Database Lookup of CRGs

To reduce the number of necessary ANI calculations a kmer profile-based lookup of CRGs against custom databases is carried out. This step is implemented via Mash parameterized with a Mash distance of 0.1, which was shown to correlate well with an ANI of roughly 90% (Ondov et al., 2016) and thereby establishing a lower limit for reasonably related genomes. The resulting set of CRGs is subsequently reduced to a configurable number of CRGs (default=100) with the lowest Mash distances.

Determination of RG

Mash distances used for the preliminary selection of CRGs were shown to correlate well with ANI values capturing nucleotide-level sequence similarities. However, Mash distances do not correlate well with the conDNA statistic, which captures the query sequence coverage within a certain reference sequence (Figure 1). In order to precisely calculate sequence similarities beyond the capability of kmer fingerprints and to assure that RGs share an adequate portion of the query genome, ReferenceSeeker calculates both ANI and conDNA to derive a highly specific measure of microbial genome relationships (Goris et al., 2007).

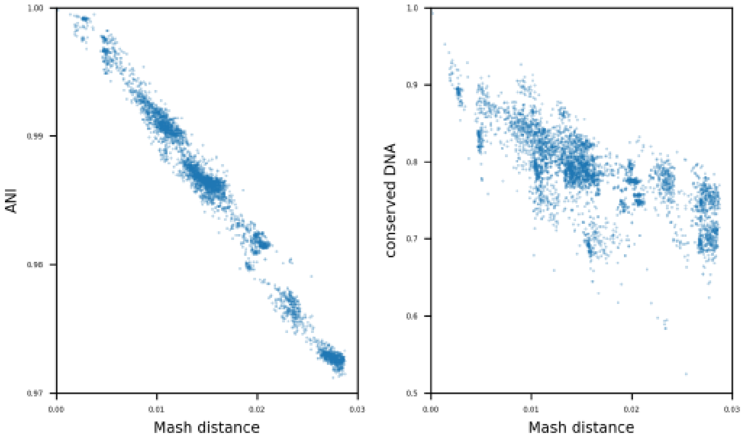
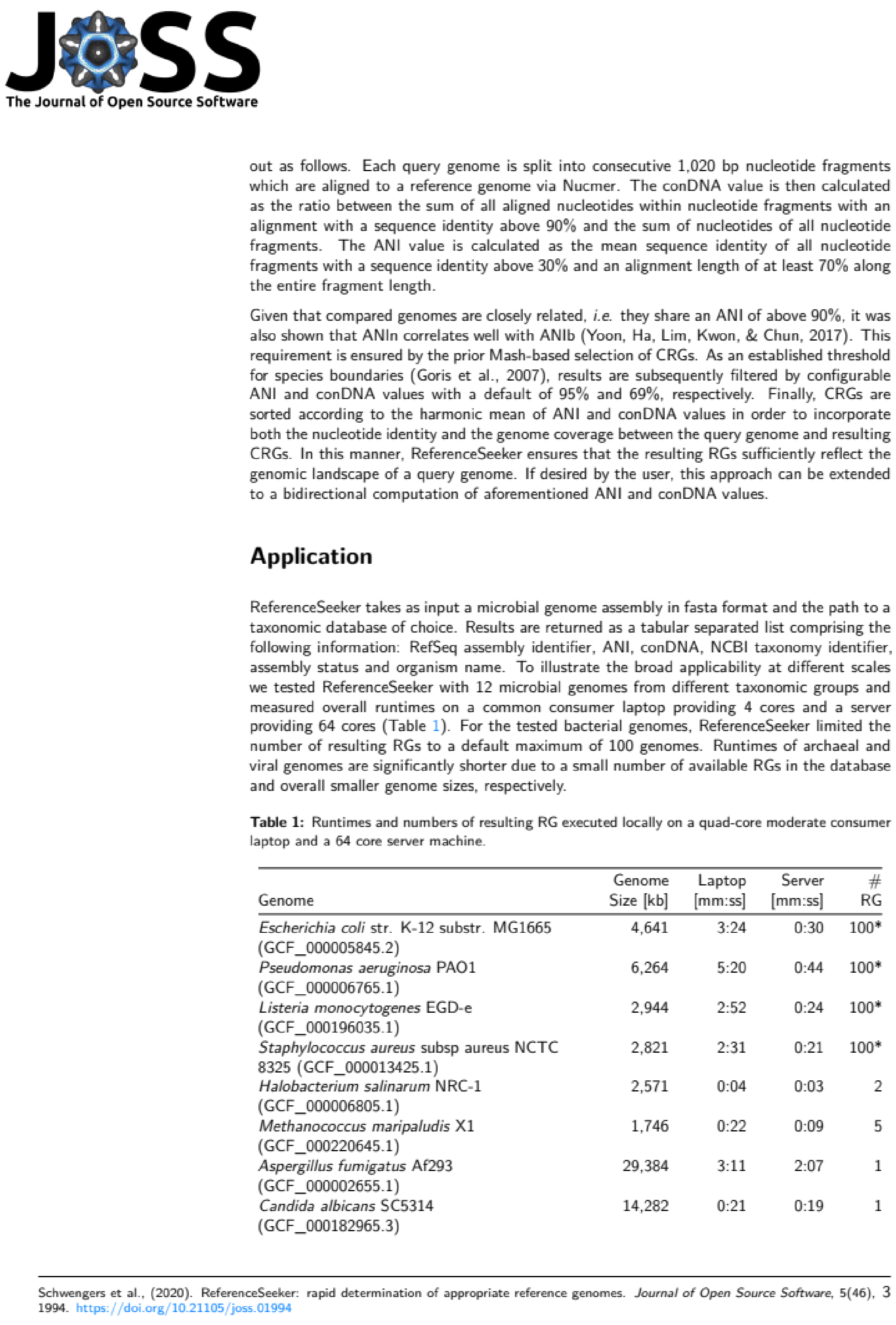


Figure 1: Figure 1: Scatter plots showing the correlation between Mash distance, ANI and conDNA values. ANI and conserved DNA values are plotted against Mash distance values for 500 candidate reference genomes with the lowest Mash distance within the bacterial database for 10 randomly selected *Escherichia coli* genomes from the RefSeq database, each.

Therefore, required sequence alignments are conducted via Nucmer of the MUMmer package (Marçais et al., 2018) as it was recently shown that Nucmer based implementations (ANIn) compare favourably to BLAST+ based implementations (ANIb) in terms of runtime. Exact calculations of ANI and conDNA values were adopted from (Goris et al., 2007) and are carried



out as follows. Each query genome is split into consecutive 1,020 bp nucleotide fragments which are aligned to a reference genome via Nucmer. The conDNA value is then calculated as the ratio between the sum of all aligned nucleotides within nucleotide fragments with an alignment with a sequence identity above 90% and the sum of nucleotides of all nucleotide fragments. The ANI value is calculated as the mean sequence identity of all nucleotide fragments with a sequence identity above 30% and an alignment length of at least 70% along the entire fragment length.

Given that compared genomes are closely related, *i.e.* they share an ANI of above 90%, it was also shown that ANIn correlates well with ANIb (Yoon, Ha, Lim, Kwon, & Chun, 2017). This requirement is ensured by the prior Mash-based selection of CRGs. As an established threshold for species boundaries (Goris et al., 2007), results are subsequently filtered by configurable ANI and conDNA values with a default of 95% and 69%, respectively. Finally, CRGs are sorted according to the harmonic mean of ANI and conDNA values in order to incorporate both the nucleotide identity and the genome coverage between the query genome and resulting CRGs. In this manner, ReferenceSeeker ensures that the resulting RGs sufficiently reflect the genomic landscape of a query genome. If desired by the user, this approach can be extended to a bidirectional computation of aforementioned ANI and conDNA values.

Application

ReferenceSeeker takes as input a microbial genome assembly in fasta format and the path to a taxonomic database of choice. Results are returned as a tabular separated list comprising the following information: RefSeq assembly identifier, ANI, conDNA, NCBI taxonomy identifier, assembly status and organism name. To illustrate the broad applicability at different scales we tested ReferenceSeeker with 12 microbial genomes from different taxonomic groups and measured overall runtimes on a common consumer laptop providing 4 cores and a server providing 64 cores (Table 1). For the tested bacterial genomes, ReferenceSeeker limited the number of resulting RGs to a default maximum of 100 genomes. Runtimes of archaeal and viral genomes are significantly shorter due to a small number of available RGs in the database and overall smaller genome sizes, respectively.

Table 1: Runtimes and numbers of resulting RG executed locally on a quad-core moderate consumer laptop and a 64 core server machine.

Genome	Genome Size [kb]	Laptop [mm:ss]	Server [mm:ss]	# RG
Escherichia coli str. K-12 substr. MG1665 (GCF_000005845.2)	4,641	3:24	0:30	100*
Pseudomonas aeruginosa PAO1 (GCF_000006765.1)	6,264	5:20	0:44	100*
Listeria monocytogenes EGD-e (GCF_000196035.1)	2,944	2:52	0:24	100*
Staphylococcus aureus subsp aureus NCTC 8325 (GCF_000013425.1)	2,821	2:31	0:21	100*
Halobacterium salinarum NRC-1 (GCF_000006805.1)	2,571	0:04	0:03	2
Methanococcus maripaludis X1 (GCF_000220645.1)	1,746	0:22	0:09	5
Aspergillus fumigatus Af293 (GCF_000002655.1)	29,384	3:11	2:07	1
Candida albicans SC5314 (GCF_000182965.3)	14,282	0:21	0:19	1

The Journal of Open Source Software

Genome	Genome Size [kb]	Laptop [mm:ss]	Server [mm:ss]	# RG
Entamoeba histolytica HM-1:IMSS (GCF_000208925.1)	20,835	6:04	4:41	1
Plasmodium falciparum 3D7 (GCF_000002765.4)	23,326	2:52	1:49	1
Influenza A virus (GCF_001343785.1)	13	0:03	0:02	1
Human coronavirus NL63 (GCF_000853865.1)	27	0:04	0:02	1

Availability

The source code is available on GitHub under a GPL3 license: https://github.com/oschwengers/referenceseeker. The software is packaged and publicly available via BioConda. Pre-built databases for bacteria, archaea, fungi, protozoa and viruses are hosted at Zenodo: https://doi.org/10.5281/zenodo.3562005. All installation instructions, examples and download links are provided on GitHub.

Funding

This work was supported by the German Center of Infection Research (DZIF) [DZIF grant 8000 701–3 (HZI), TI06.001 and 8032808811 to T.C.]; the German Network for Bioinformatics Infrastructure (de.NBI) [BMBF grant FKZ 031A533B to A.G.]; and the German Research Foundation (DFG) [SFB-TR84 project A04 (TRR84/3 2018) to T.C., KFO309 Z1 (GO 2037/5-1) to A.G., SFB-TR84 project B08 (TRR84/3 2018) to T.H., SFB1021 Z02 (SFB 1021/2 2017) to T.H., KFO309 Z1 (HA 5225/1-1) to T.H.].

Authors declare that there are no conflicts of interest.

Acknowledgement

The authors thank Karina Brinkrolf for valuable discussions, testing and bug reports.

References

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, *35*(4), 316–319. doi:10.1038/nbt.3820

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, *57*(1), 81–91. doi:10.1099/ijs.0.64483-0

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., et al. (2018). RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, *46*(D1), D851–D860. doi:10.1093/nar/gkx1068

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944. doi:10.1371/journal.pcbi.1005944

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, *14*, 60. doi:10.1186/1471-2105-14-60

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using minhash. *Genome Biology*, *17*(1), 132. doi:10.1186/s13059-016-0997-x

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *110*(10), 1281–1286. doi:10.1007/s10482-017-0844-4

8 Further contributions

Shared first authorships in peer-reviewed publications

ReadXplorer 2-detailed read mapping analysis and visualization from one single source

Rolf Hilker & Kai Stadermann & **Oliver Schwengers**, Evgeny Anisiforov, Sebastian Jaenicke, Bernd Weisshaar, Tobias Zimmermann, Alexander Goesmann

Bioinformatics, 2016

Coauthorships in peer-reviewed publications

Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning

Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, **Oliver Schwengers**, Dominik Heider

Bioinformatics, 2021

Acinetobacter stercoris sp. nov. isolated from output source of a mesophilic german biogas plant with anaerobic operating conditions

Dipen Pulami, Thorsten Schauss, Tobias Eisenberg, Jochen Blom, **Oliver Schwengers**, Jennifer K Bender, Gottfried Wilharm, Peter Kämpfer, Stefanie P Glaeser

Antonie van Leeuwenhoek, 2021

Detection of blaCTX-M-27-encoding *Escherichia coli* ST206 in Nigerian poultry stocks

Funmilola A Ayeni, Jane Falgenhauer, Judith Schmiedel, **Oliver Schwengers**, Trinad Chakraborty, Linda Falgenhauer

Journal of Antimicrobial Chemotherapy, 2020

Multidrug-Resistant and Clinically Relevant Gram-Negative Bacteria Are Present in German Surface Waters

Linda Falgenhauer, **Oliver Schwengers**, Judith Schmiedel, Christian Baars, Oda Lambrecht, Stefanie Heß, Thomas U. Berendonk, Jane Falgenhauer, Trinad Chakraborty and Can Imirzalioglu

Frontiers in Microbiology, 2019

Whole-Genome Sequences of Clinical *Enterobacter bugandensis* Isolates from Germany

Jane Falgenhauer, Can Imirzalioglu, Linda Falgenhauer, Yancheng Yao, Anja M. Hauri, Beate Erath, **Oliver Schwengers**, Alexander Goesmann, Harald Seifert, Trinad Chakraborty, Swapnil Doijada

Microbiology Resource Announcements, 2019

Detection and Characterization of ESBL-Producing *Escherichia coli* From Humans and Poultry in Ghana

Linda Falgenhauer, Can Imirzalioglu, Kwabena Oppong, Charity Wiafe Akenten, Benedikt Hogan, Ralf Krumkamp, Sven Poppert, Vinzent Levermann, **Oliver Schwengers**, Nimako Sarpong, Ellis Owusu-Dabo, Jürgen May and Daniel Eibach

Frontiers in Microbiology, 2019

Bordetella pseudohinzii targets cilia and impairs tracheal cilia-driven transport in naturally acquired infection in mice

Alexander Perniss, Nadine Schmidt, Corinne Gurtner, Kristina Dietert, **Oliver Schwengers**, Markus Weigel, Julia Hempe, Christa Ewers, Uwe Pfeil, Ulrich Gärtner, Achim D. Gruber, Torsten Hain & Wolfgang Kummer

Scientific Reports, 2018

The Regulatory Small RNA MarS Supports Virulence of *Streptococcus pyogenes*

Roberto Pappesch, Philipp Warnke, Stefan Mikkat, Jana Normann, Aleksandra Wisniewska-Kucper, Franziska Huschka, Maja Wittmann, Afsaneh Khani, **Oliver Schwengers**, Sonja Oehmcke-Hecht, Torsten Hain, Bernd Kreikemeyer & Nadja Patenge

Scientific Reports, 2017

A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies

Jia Yu, Jochen Blom, Stefanie Glaeser, Sebastian Jaenicke, Tobias Juhre, Oliver Rupp, **Oliver Schwengers**, Sebastian Spänig, Alexander Goesmann

Journal of Biotechnology, 2017

Scientific conferences

Oral presentation, VAAM 2021, online, 2021:
Microbial bioinformatics – Applications from bed to space

Oral presentation, DGH / VAAM, 6th Joint Conference of the DGHM & VAAM, Germany, Leipzig, 2020:
ASA^{3}P – an automatic and scalable pipeline for the assembly, annotation and higher level analysis of closely related bacterial isolates

Poster presentation, ASM, Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines, Washington, USA, 2018:
ASA^{3}P: An automatic and highly scalable pipeline for bacterial genome assembly, annotation and higher-level analyses

Poster presentation, DGI / DZIF, Joint Annual Meeting, Hamburg, Germany, 2017:
ASA^{3}P: An automatic and highly scalable pipeline for bacterial genome assembly, annotation and higher-level analyses

Poster presentation, DZIF, Annual Meeting, Cologne, Germany, 2016:
ASA^{3}P: An automatic and highly scalable pipeline for bacterial genome assembly, annotation and higher-level analyses

Supervised theses

M.Sc., Jan Christoph Keller, 2021: Expansion and evaluation of the software tool ReferenceSeeker towards the determination of reference genomes for bacterial cohort studies

B.Sc., Noah Elija Knoppik, 2020: Enhancement of an existing short-read assembly workflow by read-based draft assembly polishing in ASA^{3}P

M.Sc., Fabrice Hess, 2019: Design and implementation of a rapid approach for the calculation of bacterial pan-genomes based on bidirectional best blast scores

M.Sc., Maike Carina Weber, 2019: Development of a machine learning framework for microbial phenotype prediction based on protein sequences

M.Sc., Ulrich Purath, 2018: ASA^{3}P-Annotator: Ein Tool zur genomischen Annotation von Prokaryoten auf der Basis von Proteinclustern

M.Sc., David Parzych, 2018: Assessment and implementation of a bacterial analysis workflow for clinical NanoPore MinION sequence data running on minimal compute resources

M.Sc., Patrick Barth, 2017: Implementation and benchmarking of a new method extracting plasmid contigs from draft assemblies

M.Sc., Andreas Hoek, 2017: Scalable analysis of bacterial genomes in the cloud with ASA^{3}P

M.Sc., Saba Nassir, 2017: Phylogenomische und komparative Genomanalyse von *Arthrobacter mysorens* DSM28748

B.Sc., Hermann Finke, 2017: Implementation of a web-based user interface for the dynamic presentation of protein sequence annotation results.

9 Abbreviations

ANI	average nucleotide identity
AMR	antimicrobial resistance
AST	antimicrobial susceptibility testing
CCI	cloud computing infrastructure
CDC	Centers for Disease Control and Prevention
CPU	central processing unit
DNA	deoxyribonucleic acid
DZIF	German Center for Infection Research
ECDC	European Centre for Disease Prevention and Control
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
EUCAST	European Committee on Antimicrobial Susceptibility Testing
FAIR	findability, accessibility, interoperability and reusability
FASTA	text format for DNA, RNA and protein sequences
GB	gigabyte
GWAS	genome wide association studies
HT	high-throughput
HTML	hypertext markup language
HPC	high-performance computing
IT	information technology
MLST	multilocus sequence typing
MPS	marker protein sequence
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
ONT	Oxford Nanopore Technologies

PFGE	pulsed-field gel electrophoresis
RAM	random access memory
RDS	replicon distribution score
RNA	ribonucleic acid
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
WGS	whole-genome sequencing
WHO	World Health Organization

10 Acknowledgements

I’d like to take this chance to thank all the inspiring and exceptional people who supported me during the last years in regard to this thesis and made it possible in one way or another. For sure, the following list is by no means complete, and many more deserved to be mentioned. I was very lucky to make these acquaintances and I remain truly grateful for all of them. Besides these extraordinary people, I am also indebted to the German Federal Ministry of Education and Research (BMBF), the German Center for Infection Research (DZIF) and the German Network for Bioinformatics Infrastructure (de.NBI) for providing the funding for me and my research.

Alex, It’s been such a long way … and it all began in a non-existing town (so they say) with a rough idea that was accepted and supported by you for my master thesis and later on happened to result in the beginning of my PhD journey. It was a fascinating time with countless things to learn and unexpected turns. I’d like to thank you for all your continuous support and constructive feedback. For sure, some challenging times were included, as well, but over the course of all these years, you consistently provided me with the required resources, trust and generous freedom in terms of research ideas and, in particular, physical presence and time. This has never been taken for granted – a huge thank you!

Prof. Dr. Trinad Chakraborty, thank you so much for opening the door to the exciting and fascinating field of WGS-based medical microbiology!

Torsten, I am very grateful for your day-to-day professional help and guidance. You facilitated my first steps in medical microbiology and helped me to find my way around this unknown terrain. Last but not least, I’d like to thank you very much for your feedback and all your help in all respects.

Bioinformatics & **medical microbiology colleagues**, a huge shoutout to all current and former colleagues in the Computational Bio group, the Institute of Medical Microbiology and the Computational Genomics group. Thanks a lot for all the help, support, interesting and fruitful discussions and last but not least all the fun! In particular, I’d like to thank **Sebastian** (Jünemann) for his technical help and support in the very early days, slowly transitioning myself from a trained biologist into an applied microbial bioinformatician. He endured a lot. I’d like to thank **Jochen** for tons of inspiring discussions, feedback, help, beneficial little pushes every now and then, and last but not least also for all the fun nights in Bielefeld and Giessen! I’d like to thank **Sebastian** and **Karina** for their feedback, comments and final corrections and Sebastian also for his enduring technical support over the years.

My entire **family**, where to start? I wholeheartedly thank you for everything ever since. Thanks for all the trust and loving support that enabled me to always pursue my interests and to make my way.

And, in particular, **Anne**, **Lasse** & **Peri**, from the bottom of my heart, thank you for all your support, forbearance and patience. And for everything beyond that – without words.

11 Declaration

I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and cited all text passages that are derived literally from or are based on the content of published work of others, and all information relating to verbal communications. I consent to the use of an anti-plagiarism software to check my thesis. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ in carrying out the investigations described in this dissertation.”

Giessen, December 2021
Oliver Schwengers