Bioinformatic analysis of the microbial metagenome in cystic fibrosis airways.

A thesis submitted for the degree of

Doctor of Philosophy (PhD)

in the subject of Bioinformatics

by

Patricia Morán Losada, M.Sc Bioinformatics

November 2015

International PhD program "Infection Biology"

Hannover Medical School,

Clinic for Paediatric Pneumology and Neonatology

PhD Project funded by the Bundesministerium für Forschung und Technologie, German Center for Lung Research (DZL), and the Mukoviszidose e.V. (project 1206)

Acknowledged by the PhD committee and head of Hannover Medical School

President:	**Prof. Dr. Christopher Baum**
Supervisor:	**Prof. Dr. med. | Dr. rer. nat. Burkhard Tümmler, Clinic for Paediatric Pneumology and Neonatology, Hannover Medical School**
Co-Supervisor:	**Dr. Colin Davenport; Prof. Dr. Alexander Goesmann, Justus-Liebig-University Giesen**
External expert:	**Prof. Dr. Trinad Chakraborty, Institut für Medizinische Mikrobiologie, Universität Giessen.**
Internal expert:	**Prof. Dr. Susanne Häussler, Molecular Bacteriology, Helmholtz Centre for Infection Research**

Day of public defence: 22nd January 2016

“Science and everyday life cannot and should not be separated”

Rosalind Franklin

“DNA is like a computer program but far, far more advanced than any software ever created”

Bill Gates

Table of contents

Abstract

Due to the rapid advances in sequencing technologies, large amounts of DNA sequences can be obtained quickly and cheaply, therefore, sequencing has become a routine procedure to perform analysis of microbial organisms. Shotgun metagenomic DNA sequencing is a relatively new scientific field in which genetic material is extracted directly from the environment, without prior cultivation or amplification. Therefore, metagenomics has become a robust environmental sequencing approach which provides insight into community biodiversity and function. However, the analysis of metagenomic sequences requires specific computational tools due to the biases and errors associated with the high-throughput sequencing technology as well as the complex structure of the data.

This thesis focuses on the metagenomic analysis of samples collected from individuals with cystic fibrosis (CF). The dissertation presents the first unbiased and exhaustive metagenomic analysis of cystic fibrosis sputum samples to date. Relative and absolute abundances of DNA viruses, bacteria and fungi were calculated which demonstrate that a large repertoire of microbial organisms are present in the CF lower airways. On average, several hundred of bacterial taxa which made up more than 99% of the microbial community were identified in the analysis.

A second observation is that each individual carries a specific microbial signature of multiple lowly abundant species superimposed by few disease-associated pathogens such as *P. aeruginosa* and *S. aureus*. Furthermore, the analysis of the three age-group individuals (children, adolescents and adults) indicates that the CF microbiome of our cohort is more diverse in children, which involves a healthier state of the individual. This diversity is lost as the age of the individual advances. i.e. the older a patient gets, they become dominated by one or two specific pathogens. However, there are a few cases of pancreatic sufficient (PS) CF adults which were found carrying a healthy microbial metagenome. Anaerobes were also identified in higher proportions in young individuals and their proportion decreases with age.

Previous studies have addressed that the *S. aureus* and *P. aeruginosa* populations in CF individuals only consisted of between one and three major clones. However, a key finding of this thesis identifies several clones of *S. aureus* and *P. aeruginosa* present within the microbial community of an individual. In addition, the identification of the major clones of *S. aureus* and *P. aeruginosa* was performed using the MLST database and multi-marker array, respectively. Due to the low sequencing coverage present in some samples, just a few strains were identified. Four out of ten *P. aeruginosa* strains belonged to ubiquitous clones in the global *P. aeruginosa* population and two pairs out of thirteen *S. aureus* strains were assigned to the common clone type ST7 and the pandemic MRSA ST22.

As well, the identification and characterization of antibiotic resistance genes present in these two key species was performed. Mutations in the gyrase-encoding *gyr* loci were found in *S. aureus*.

To achieve a high accuracy output, a new model was implemented in the metagenomic analysis. This model reduces the GC content biases present in SOLiD technology and normalizes the data based on the genome length.

A second objective of this dissertation was the study of bacterial recombination, specifically in the dominant cystic fibrosis pathogens *P. aeruginosa* and *S. aureus*. It is already known that pathogenic and non-pathogenic bacteria can belong to the same taxa having similar DNA sequences. Therefore, the analysis of recombination is needed to understand the phylogenetic relationship between taxa. A new algorithm to study recombination based on the haplotype reconstruction was developed and applied to study the variation present in the previous mentioned dominant pathogens. The median haplotype length found was 51bp for *S. aureus* and 99bp for the unrelated *P. aeruginosa* clones. However, the intraclonal analysis of *P. aeruginosa* showed that haplotypes are 1000-fold longer within clone than among unrelated clones.

Acknowledgments

I have just completed three years of my thesis and when I look back on my experience I see a lot of happy memories and also plenty of relief that I have finished it. I'm sure I've experienced a similar range of emotions like all of the other PhD students.
This PhD has been a huge challenge for me, especially when you have to speak a language which is not your mother tongue or the mother tongue of the people involved. All of these memories will always bring a smile to my face in the future.

First of all, I would like to give an enormous thank to my supervisor Burkhard Tümmler, you have been a tremendous mentor for me, even if some times I didn't know the maths or statistics from the "high school". You have been very patient when you have to explain the same things to me over and over again. THANK YOU.
Thanks to Philippe Chouvarine, the person who taught me the immense value of Google to have a successful bioinformatics PhD, always saying: "that it is easy, just search in google". But he didn't know how annoying I can be in the pursuit to find a human answer. Thanks.

Thanks to Sebastian Fisher for his continuous support and great humor everyday in the office. Thanks to Lutz Wiehlmann, the guy who is always running from one side to the other doing a lot of things at the same time, but always finding time to support all of his colleagues.

Huge thanks to my colleagues from the wet-lab Silke Hedtfeld, Samira Mielke, Angela Schulz and Marie Dorda, without whom I could not have done this project.
Thanks to Helga Riehn-Kopp for all administrative work, Jens Klockgether for all biology and photoshop support. Thanks also to Nina Cramer, Antje Munder, Sarah Dethlefsen, Stephanie Tamm, Chidiebere Awah, Frauke Stanke for listening to my bioinformatic talks and support.

I would also like to thank Professor Dr. Trinad Chakraborty, Professor Dr. Susanne Häußler and Dr. Colin Davenport for serving as my committee members. Thanks to my co-supervisor Dr. Alexander Goesmann.

Additionally, I would like to thank my parents and my brother for always being there and listening to my scientific explanations even if their professions are so different to mine. Thanks to the rest of my family and to all my friends for their patience and support.

And last but not least, special thanks to Darius for his help with this thesis, trying to understand Science, and the meaning of haplotypes, SNPs, *Pseudomonas aeruginosa*, and all of the other complex scientific words.

Chapter 1

Thesis Outline and Research Objectives

1.1 Research Objectives

Motivation

I would like to start my dissertation with a citation from Stephen Jay Gould: "*We live now in the "Age of Bacteria." Our planet has always been in the "Age of Bacteria," ever since the first fossils—bacteria, of course—were entombed in rocks more than 3 billion years ago. On any possible, reasonable or fair criterion, bacteria are—and always have been—the dominant forms of life on Earth*."

Microorganisms are known to be the most diverse and abundant type of organisms on Earth[1-2]. They are essential for all types of life on Earth. Indeed they are involved in the cycles of carbon, nitrogen, oxygen and sulfur. Hosts depend on the associated microbial communities for obtaining necessary vitamins, nutrients and metals[3-5]. It has been shown that microbes associated with the gut of human beings enable the extraction of energy from food. This energy would not be accessible without them[6-7]. However, perturbations in the composition or functions of the microbiota can lead to metabolic and inflammatory disorders in their host[8-9]. Therefore, understanding of the complexity of a microbial community is more relevant than studying a single species of such community.

Advances in DNA sequencing technologies and bioinformatics tools allow the exploration of uncultured host-associated microbial communities. Sequencing of the 16S rRNA gene is the most widely used approach for characterizing the microbiota and its diversity[10-13]. However, whole-genome shotgun (WGS) sequencing is an alternative culture-independent method to obtain a deeper and more robust analysis of the microbial communities[14-16].

Here, we want to move one step further and employ these technological advances for the analysis of the lower airways microbiome of individuals with cystic fibrosis using whole genome sequencing.

To my knowledge, only a few studies have been conducted on the cystic fibrosis metagenome where samples from at least ten CF adults have been analyzed[17-19].

Main aims and objectives

The overall aim of this thesis was to resolve the metagenome of 25 individuals with CF, therefore temporal series of sputa samples were collected.

Specifically:

- To conduct an unbiased and comprehensive study about the frequency and abundance of bacteria, DNA viruses and fungi in the CF lower airways.
- To investigate possible differences between exocrine pancreas insufficient (PI) and exocrine pancreas sufficient (PS) CF individuals and age groups (A - children, B - adolescents and C - adults).
- To identify the major clones of the dominant CF pathogens *Pseudomonas aeruginosa* (*P. aeruginosa*) and *Staphylococcus aureus* (*S. aureus*).
- To detect antibiotic resistance genes associated with *P. aeruginosa* and *S. aureus*.
- To study the role of recombination in *P. aeruginosa* and *S. aureus*.

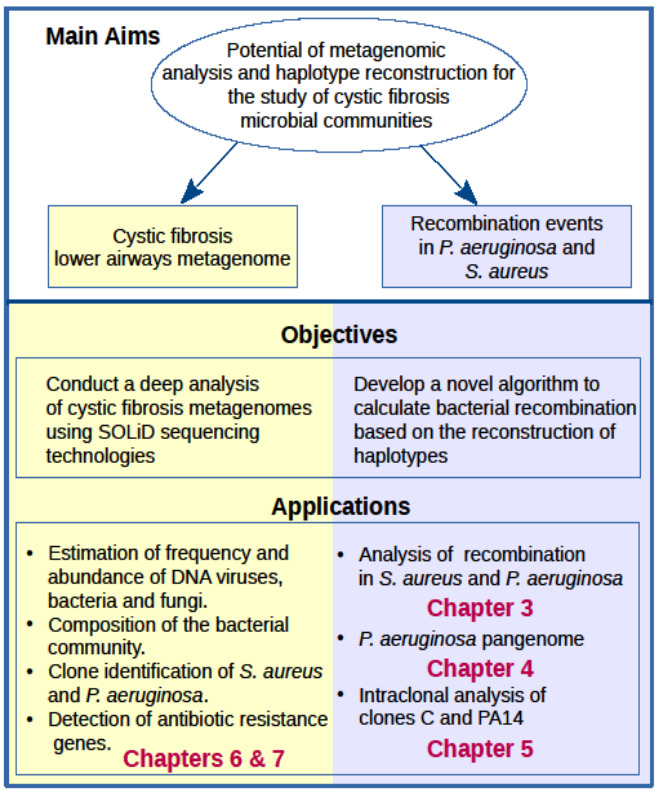


Figure 1. The figure summarizes the main objectives and applications of my dissertation and indicates the respective chapters where they are addressed.

1.2 Overview of Chapters

The following section provides an overview of the thesis with a short description of each individual chapter.

Chapter 1 outlines my PhD project and its objectives. The project consists of two main parts, first, the cystic fibrosis lower airways microbial metagenome and secondly the connected study of bacterial recombination based on the haplotype reconstruction. The chapter also provides an overview of the dissertation and concludes with a list of publications carried out during the PhD.

Chapter 2 provides an introduction to metagenomics and its advantages over other technologies followed by an overview of both next generation sequencing technologies and third generation sequencing. The chapter continues by explaining cystic fibrosis and *P. aeruginosa*. It concludes with an explanation of bacterial recombination.

Chapter 3 presents a novel algorithm to study bacterial recombination based on the reconstruction of haplotypes and its application to *S. aureus*.

Chapter 4 explores the *P. aeruginosa* pangenome where my main contribution was to apply the aforementioned algorithm and to study interclonal recombination.

Chapter 5 identifies the intraclonal genome diversity of the major *P. aeruginosa* clones C and PA14.

Chapter 6 highlights the challenges to perform an accurate metagenomic analysis and the biases found in SOLiD technologies. It outlines a new bioinformatic model to filter and normalize metagenome analysis.

Chapter 7 presents the main focus of the dissertation, i.e. the execution of the largest and exhaustive study to date of cystic fibrosis lower airways metagenomes.

Chapter 8 is the final chapter where I present the conclusion of the dissertation and future perspectives.

Chapter 9 contains the bibliography.

Chapter 10 provides supplementary information regarding the metagenomics pipeline used in the analysis and the haplotype reconstruction algorithm.

1.3 Overview of publications

- **Losada PM**, Chouvarine P and Tümmler B. Bacterial recombination analysis based on haplotype construction. [Submitted].
- Hilker R, Munder A, Klockgether J, **Losada PM**, Chouvarine P, Cramer N, Davenport CF, Dethlefsen S, Fischer S, Peng H, Schönfelder T, Türk O, Wiehlmann L, Wölbeling F, Gulbins E, Goesmann A, Tümmler B. Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment. [Environ Microbiol, 2015].
- Fischer S, Cramer N, **Losada PM**, Chouvarine P, Dethlefsen S, Davenport C, Dorda M, Goesmann A, Hilker R, Mielke S, Schönfelder T, Suerbaum S, Türk O, Woltemate S, Wiehlmann L, Klockgether J and Tümmler B. Intraclonal genome diversity of the major *Pseudomonas aeruginosa* clones C and PA14. [Environmental Microbiology: in revision].
- Chouvarine P, Wiehlmann L, **Losada PM** and Tümmler B. Filtration and normalization of sequencing read data in whole-metagenome shotgun samples. [Submitted].
- **Losada PM**, Chouvarine P, Dorda M, Hedtfeld S, Mielke S, Schulz A, Wiehlmann L, Tümmler B. The cystic fibrosis lower airways microbial metagenome. [Submitted].

Conference Publications & Talks

Losada PM. Haplotyping the *Pseudomonas aeruginosa* pangenome. Cell Symposia Microbiome & Host Health. Lisbon 2013; and First International Symposium on Paediatric Research. Hannover 2013. [Poster].

Losada PM, Chouvarine P, Dorda M, Hedtfeld S, Mielke S, Schulz A, Wiehlmann L, Tümmler B. The cystic fibrosis lower airways microbial metagenome. 9th European CF Young Investigator Meeting. Paris 2015. [Poster & Talk].

Losada PM, Chouvarine P and Tümmler B. Metagenomic and genomic haplotype analyses of microbial communities. 10th CeBiTec Symposium. Bielefeld 2015. [Poster].

Losada PM, Chouvarine P, Dorda M, Hedtfeld S, Mielke S, Schulz A, Wiehlmann L, Tümmler B. The cystic fibrosis lower airways microbial metagenome. DZL annual meeting. Hamburg 2015; EMBL Conference on the Human Microbiome. Heidelberg 2015; and 6th European Conference on Prokaryotic and Fungal Genomics. Göttingen 2015. [Poster].

Chapter 2

Background

2.1 Metagenomics

In traditional genomics, cultivation of microbes is required to sequence new microorganisms. However, it has been shown that culture-dependent methods have several limitations because in some environments there are microbes that cannot be cultured[20], and, therefore, it is impossible to calculate the relative abundance of species present in a specific habitat.
With the ongoing advances in sequencing technologies, especially with the arrival of next generation sequencing (NGS) during the last decade, the study of microbial communities as a whole has become possible[21-22]. Therefore, metagenomics offers a new path to study microbial communities, their species diversity, structures, phylogenetic composition, metabolic and functional diversity.
Metagenomics is the study of microbial communities, directly from their natural environment, without prior cultivation of single organisms and amplification of their DNA[23-25]. The term metagenomics was first coined in 1998 by Handelsman *et al.*[26] in the context of "analysis of the collective genomes of soil microflora".

It is important to note that the term metagenomics refers to the shotgun sequencing where all DNA from the sample is sequenced without amplification[27-28].

In 2004 the first large scale shotgun metagenome study[1] was conducted, starting a new era in this field. In the following years, the low costs of sequencing and development of 454 pyrosequencing had facilitated two large metagenome projects, the Metagenomics of the Human Intestinal Tract (MetaHIT)[29] and the American Human Microbiome Project (HMP)[10]. To date, an immense number of additional metagenomes have been sequenced. All of the data is available in public resources like MG-RAST[30], IMG[31], GOLD[32] and NCBI[33].

2.1.1 16S rDNA microbial sequencing

The 16S rDNA gene encodes the 16S rRNA which is a component of the prokaryotic ribosomes. It consists of highly conserved and variable regions (Figure 2).

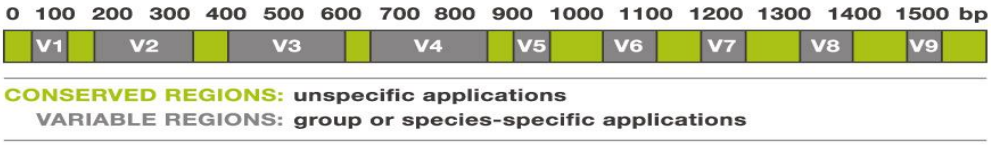


Figure 2. 16S rDNA gene showing variable and conserved regions.

Woese *et al*.[34-35] have shown that phylogenetic relationships among bacteria and archaea can be determined by comparing ribosomal rDNA genes. PCR of bacterial DNA with "universal" primers that are complementary to conserved regions yield sequences with species-specific signatures. Therefore 16S rDNA sequencing can provide insight into the diversity and structure of a microbial community[36-37].

While powerful, 16S rDNA sequencing is not without limitations:

- It has biases associated with the PCR amplification step which may result in failure to identify all members of the bacterial community[38-40].
- Inaccurate diversity estimations could be produced because of different capability of resolving taxa[41-42].
- It cannot resolve the biological functions associated with the taxa present in the community.

These limitations make metagenomics a more suitable technology to achieve all objectives of this dissertation. However, since the sequencing technologies that are used in this thesis produce short single-end reads, taxonomic analysis was more difficult than what it could be if we used a sequencing platform producing longer paired-end reads.

2.1.2 Computational Challenges in Metagenomics

Metagenomic analysis is facing various computational problems such as genome assembly or taxonomic classification. Several review papers address these challenges[22,23,27,43].

- Genome assembly: Assembling short sequences becomes extremely difficult in metagenomics. Highly diverse microbial communities require high sequencing depth to obtain complete genomes of lowly abundant species. Even if enough sequences have been obtained, sequences from different homologous species could be assembled together producing chimeric contigs[44-45].

- Sequence classification: Determining the genome from which a read was derived is one of the primary goals of metagenome studies. Taxonomic classification methods follow one of the two approaches (i) composition based approach and (ii) the comparative approach. The first method is based on nucleotide composition (i.e. k-mer abundance) or GC-content, which is compared with features computed from reference sequences with known taxonomic origin. TETRA or PhyloPythia are some examples of k-mer frequency software tools. On the other hand, comparative approaches rely on homology obtained by conducting databases searches. Such methods perform alignments of sequences or contigs using global and local sequence algorithms that generate most matches to a reference database. In general, these methods can be divided into methods that are based on Basic Local Alignment Search Tool (BLAST)[46] homology searches and those which are based on Hidden Markov Model (HMM)[47] homology searches.

In this dissertation comparative approaches based on alignments against databases will be performed. More information regarding the approach used can be found in **Appendix1**.

2.2 Sequencing Technologies

2.2.1 First Generation Sequencing

Almost twenty years after the discovery of the double-helix structure of DNA, the first DNA sequencing was conducted by Frederick Sanger in 1975[48]. In 1977 the method was improved introducing radioactive or fluorescently labeled dideoxynucleotides (ddNTP) as chain terminators. This method can produce sequences up to 1,000bp with a low error rate. Around the same time Maxam and Gibert presented a non-enzymatic method, which involves chemical modification of the DNA sequence. However, this method produces shorter sequences (up to 100bp). High cost per base and low throughout are the main limitations of first generation sequencing.

2.2.2 Second/Next Generation Sequencing (NGS)

Lower cost and massively parallel sequencing are the main advantages of NGS. The high throughput produced by these technologies enables new research projects such as metagenomics.

454 Pyrosequencing

The 454 pyrosequencing technology is based on a principle called sequencing-by-synthesis. First, DNA is denatured and fragmented and secondly each fragment of single-stranded DNA is bound to a bead which is encapsulated into a water droplet within an oil phase for emulsion PCR (emPCR) amplification. The beads are localized in wells on a plate, and each well contains

at most one bead carrying a unique single-stranded DNA fragment. During the pyrosequencing process, light is generated that is proportional to the number of incorporated nucleotides. This step is where most errors arise, especially in homopolymeric regions of 3 or more nucleotides (Figure 3 a).

Initially, this technology had a read length of 100bp but can now produce an average read length of 400bp.

SOLiD (Sequencing by Oligonucleotide Ligation and Detection)

SOLiD employs sequencing-by-ligation (Figure 3 b). A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. After that, a di-base probe is ligated to the template DNA, the dye is cleaved off and images are captured. Then, a new cycle begins 5 bases upstream from the priming site. The process is repeated over seven cycles and the process is repeated for five rounds. In each round a new primer is hybridized offset by one base (n, n-1, n-2, n-3 and n-4) (Figure4). This technology has a read length of 75bp. The rate of accuracy is about 99.94 % over the whole sequence length due to the specificity of the di-base probe which is interrogated every 1st and 2nd base at each ligation reaction[48].

a **b**

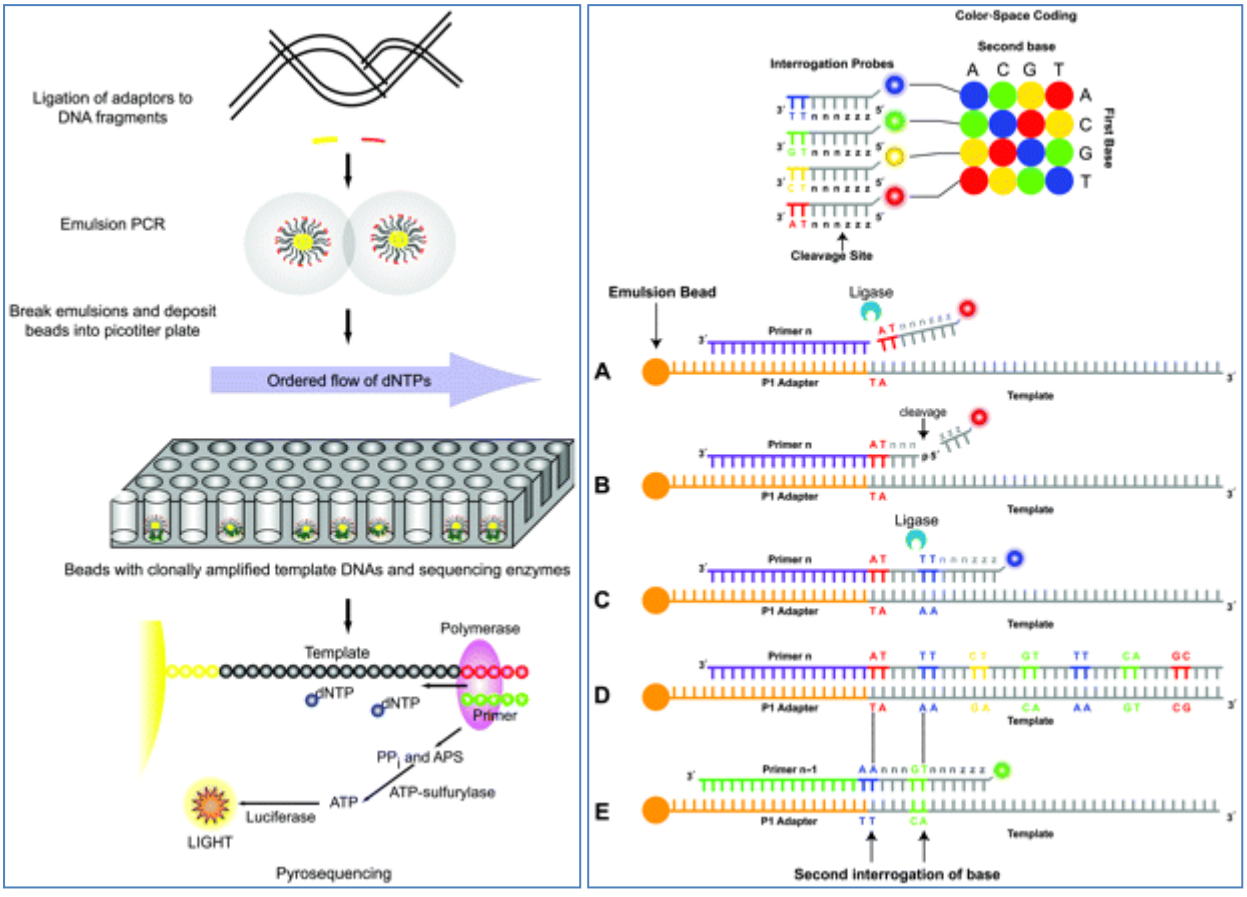


Figure3. Next-generation sequencing technologies that use emulsion PCR. **a** Pyrosequencing using Roche 454. **b** sequencing by ligation method using SOLID (Sequencing by Oligonucleotide Ligation and Detection). (Figure taken from reference 48).

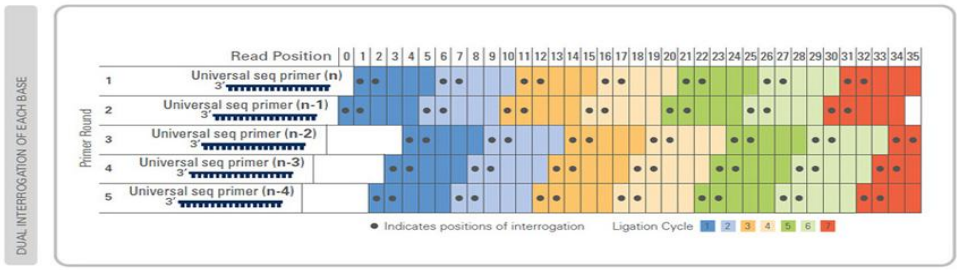
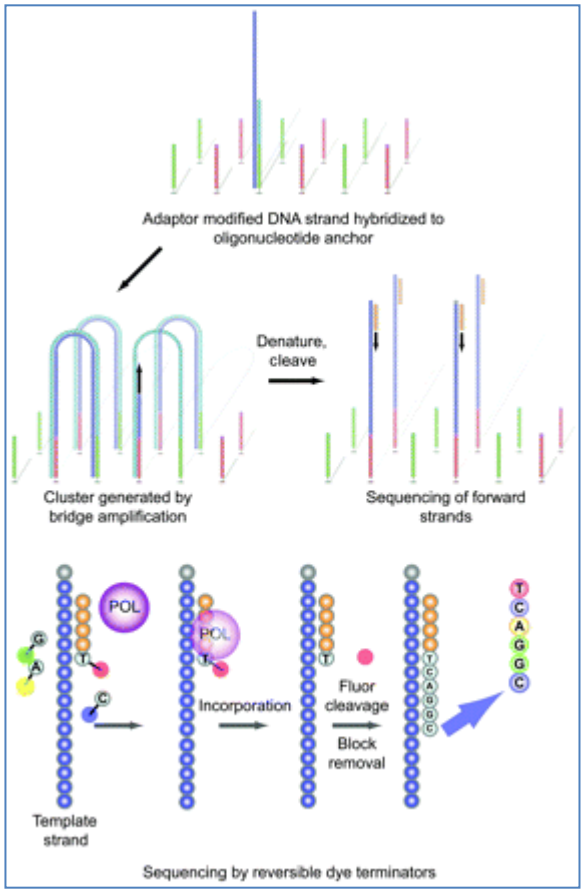


Figure 4. Dual interrogation of each base (Figure taken from Applied Biosystems website).

Illumina/Solexa

The Illumina/Solexa Genome Analyzer is the most widely used system to date. Illumina sequencing is also sequencing-by-synthesis. Sequencing templates are immobilized on a flow cell surface. During the library preparation DNA is fragmented and adapters are appended. These adapters are necessary to bind to the complementary sequencing templates of the flow cell.



After binding of the libraries to the flow cell the clustering step starts. Each attached oligo is extended and a complementary strand is synthesized. This double stranded DNA is denatured and the original fragments are washed away. Each single molecule of bound DNA is then amplified by a process called bridge PCR (bPCR). The iteration of this process produces dense clusters of copies around the initial fragment.

Next, sequencing is performed (Figure 5). Sequencing primers are added and hybridized to the immobilized fragments. Only one base with fluorescently labeled dNTPs is added in each cycle to extend these primers. The labels are excited with a laser and emitted light signals are recorded. The identification of nucleotides is based on the signal intensity.

Figure 5. Sequencing-by-synthesis of Illumina
(Figure taken from reference 48).

Ion Torrent Sequencing

Also known as pH-mediated sequencing, the method is based on the pH changes occurring in the solution. Each time a nucleotide is incorporated into a DNA strand by polymerase, a

hydrogen ion is released as a byproduct. These hydrogen ions lead to the pH change which is detected by an ion sensor. Labeling of nucleotides and imaging steps are avoided with this technique what facilitates shorter sequencing times.
This technology produces short reads of around 200 bp, it is low cost and high speed. However, error rates still remain high.

2.2.3 Third Generation Sequencing

The third generation sequencing technologies are characterized by a single-molecule real-time (SMRT) sequencing which produces longer read lengths and higher throughput. The PCR amplification step is eliminated yielding faster sequencing times.

Pacific Biosciences

Pacific Biosciences is the first third generation sequencing approach to directly observe a single molecule of DNA polymerase as it synthesizes a strand of DNA[49] without any stop to detect the incorporated nucleotide. It uses zero-mode waveguide (ZMW) technology[50]. Fluorescently labeled nucleotides are introduced in the chambers during the polymerase. The label is clipped off by the polymerase enzyme and a sensor detects the emission of the light. PacBio RS II systems can produce sequences with an average length of 8.5 kbp. However, high error rates of 11% have limited the application.

Oxford Nanopore Technologies

Oxford Nanopore involves the use of nanopores, which can be transmembrane cellular proteins or artificial holes in a silicon layer[51]. The single stranded DNA moves through the pore following an electrical field put into the chamber filled. Each nucleotide causes a specific change in the current which can be detected.

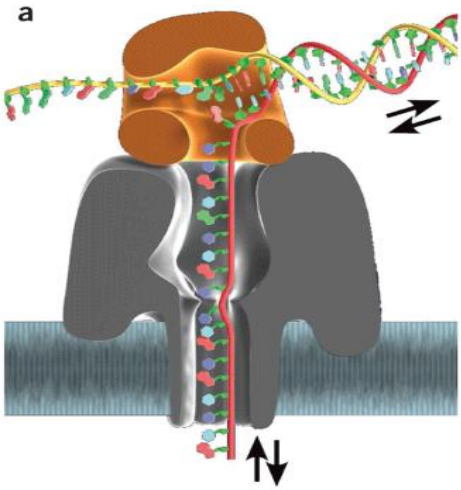


Figure 6. a) DNA inserted in a nanopore, with speed control provided by a phi29 DNA polymerase (brown) and α-hemolysin nanopore (gray). (Figure taken from 34)

2.3 Cystic Fibrosis

Cystic fibrosis (CF) is an autosomal recessive genetic disease caused by mutations of the Cystic Fibrosis Transmembrane conductance Regulator (*CFTR*) gene. It is characterized by chronic airways infections, inflammation and progressive decline in lung function. Further features are intestinal disease, pancreatic insufficiency, liver disease and CF-related diabetes. Most morbidity and mortality are determined by the progressive lung disease.

Culture-dependent methods have identified the common human pathogens, *S. aureus* and *H. influenza* as well as the opportunistic pathogen *P. aeruginosa* as the most dominant organisms associated with respiratory tract infection in CF individuals. However, many other organisms are present such as *Burkholderia cepacia complex*, *Stenotrophomonas maltophilia*, *Achromobacter* spp. and fungal pathogens such as *Candida albicans* and *Aspergillus fumigatus* (Figure 7).

Recently culture-independent methods have revealed the presence of a polymicrobial community in the airways of CF patients[52-54]. Many organisms not previously detected have been reported such as *Streptococcus* spp., *Prevotella* spp., *Veillonella* spp. and other anaerobic organisms.

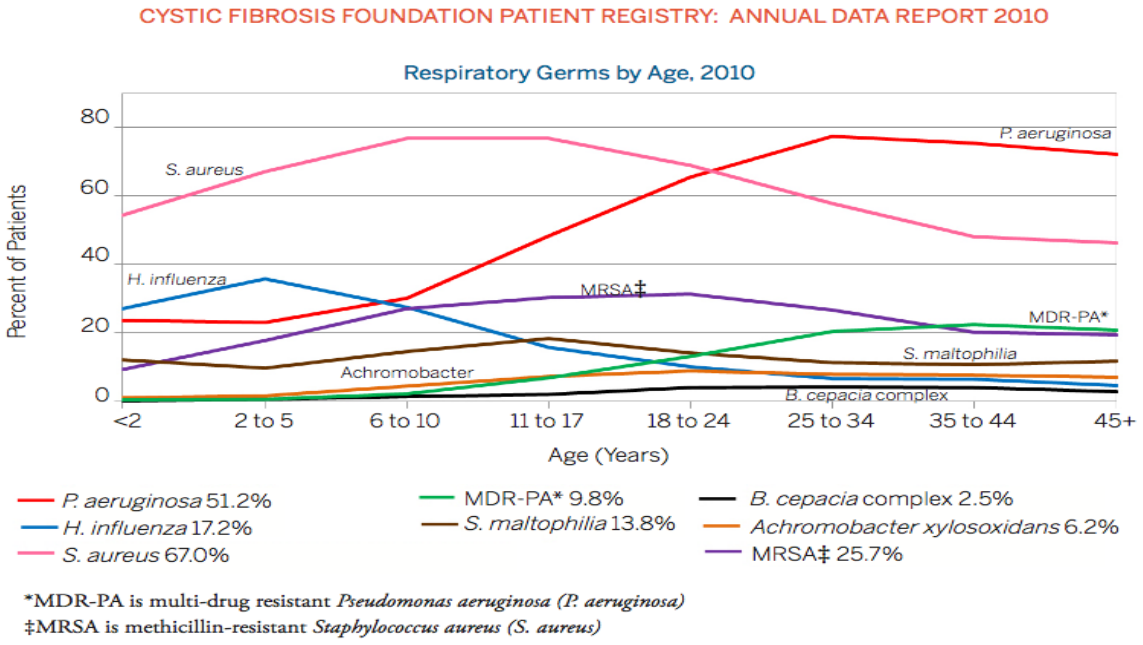


Figure 7. Prevalence of respiratory pathogens according to patients' data from Cystic Fibrosis Fundation Patient Registry, 2010 Annual Data Report.

The role of anaerobes in the progression of CF airway disease is not clear. However, recent studies have shown that anaerobes are characteristic inhabitants of the upper and lower airways of healthy non-CF humans[55] and are also associated with less inflammation and better lung function in CF individuals[56]. These results suggest that anaerobes are an indicator of health in CF patients.

The following paragraphs will highlight some characteristics of the dominant CF species with special attention to *P. aeruginosa* and *S. aureus,* which are involved in the second topic of my thesis.

Pseudomonas aeruginosa

P. aeruginosa is the most common pathogen found in CF individuals. After colonization and rapid adaptation, it established a chronic infection, which is responsible for the large majority of morbidity and subsequent mortality in CF patients[57-58]. The strong inflammatory response caused by the infection with this bacterium leads to lung damage and lung decline in CF patients.

Often, this bacterium undergoes morphological changes. It becomes immobile, reduces the production of virulence factors, changes its lipopolysaccharide (LPS) structure[59] and becomes mucoid[60]. *P. aeruginosa* has the capacity to form biofilms which contribute to colonization and chronic infections. The mucus present in the epithelial cells of CF individuals facilitates the adhesion of the biofilm to the surface. Biofilms confer multi-drug resistance and reduce the immune response of patients, what makes it more difficult to eradicate this bacterium.

This bacterium possesses surface components such as pili and flagella which are implicated in colonization and the type III secretion system which allows bacteria to inject toxins directly into the host cell.

The interclonal and intraclonal population structure of *P. aeruginosa* as well as the role of recombination in this bacterium will be described in **Chapters 3**, **4** and **5**.

Staphylococcus aureus

S. aureus is the most common gram-positive bacterial pathogen recovered from the respiratory tract of infants and children with CF. *S. aureus* causes an inflammatory response which can lead to irreversible lung damage.

This bacterium grows typically aerobically but also as facultative anaerobe. It is capable of forming biofilms[63]. It has developed resistance to methicillin through the acquisition of the *mec*A gene (MRSA).

Special aspects of this bacterium associated with chronic infection are the appearance of small colony variants (SCVs). SCVs are characterized by a slower growth rate, reduced expression of haemolysins, but high resistance rates against many antibiotics[61-62].

S. aureus expresses numerous virulence factors, which promote tissue colonization and tissue damage. An example is the Panton-Valentine leukocidin (PVL), which is a cytolytic toxin that forms pores in the membranes of leukocytes. It has been associated with severe skin infections and necrotizing pneumonia.

The role of recombination in this bacterium will be described in **Chapter 3.**

Burkholderia cepacia complex

Burkholderia cepacia complex (Bcc) is the collective name for a group of seventeen closely related bacteria which have been isolated from natural environment and human infections. Although it is not the most common bacterium, it is amongst the most virulent bacterial pathogens isolated from individuals with CF. Long-term respiratory infections with Bcc in CF patients generally induces a rapid decline in lung function and in some cases a fatal necrotizing pneumonia called "cepacia syndrome". Bcc bacteria are often resistant to most used antibiotics, therefore pulmonary colonization with Bcc is associated with a high risk of death[64].

Stenotrophomonas maltophilia and *Haemophilus influenzae*

S. maltophilia is a Gram-negative, biofilm-forming bacterium. It is an organism of low virulence, however, is an emerging multi-drug resistant opportunistic pathogen in CF individuals. Recently, it has been shown to be capable of colonizing CF airways with lower levels of lung function[65].

H. influenzae commonly infects the respiratory tract of young individuals with CF. It has been found to be capable of forming biofilms in the airway epithelium[66]. The non-encapsulated *H. influenzae* (NTHi) serotype is associated with chronic lung infections and acute exacerbations in CF patients. NTHi possesses adherence factors which may play a role in colonization and persistence in the human respiratory tract.

2.4 Bacterial recombination

Recombination is one of the main evolutionary processes that generate genome variation in species, it is a fundamental driving force in evolution, allowing different evolutionary histories to different species and between different regions of a genome.

Most of the recombination studies have been done on humans. However, many questions are still unanswered regarding the details of bacterial recombination process as for example, how many recombinations occur in every region of the genome or how recombination rates change over time.

Determining the frequency of recombination will be important to calculate the extent to which genes are exchanged within the same population or between different ones. Specific chromosomal regions of the genome have different recombination rates.

The importance of recombination in the evolution of bacterial pathogens has become a very hot topic of research. Changes in fitness or phenotype of bacterial genomes including increases or decreases in virulence or pathogenicity are known to be associated with recombinant genomes. Different comparative genome analyses of bacterial species have shown that the DNA exchanged and integrated in the genome seems to confer different pathogenicity levels in some strains of the same species[67].

There are three mechanisms of bacterial recombination, i.e. transduction, transformation and conjugation. Transduction is the process by which a virus transfers genetic material from one bacterium to another. Transformation involves the internalization and chromosomal integration of foreign DNA. Conjugation requires cell-to-cell contact to transmit DNA from donor to recipient. The capability of these mechanisms determines the extent of frequency of recombination and will be the driving force of adaptation and evolution of a bacterial species.

Next generation sequencing (NGS) technology has provided an increase in the number of sequenced genomes and has offered the opportunity to analyze the recombination events across a higher number of species genomes. Most of the bacterial recombination studies have used the multilocus sequence typing (MLST) procedure, which identifies all the single nucleotide polymorphisms (SNPs) over the sequence of seven housekeeping genes, given a specific allelic profile to each sequence type. This approach reveals the importance of recombination in the evolutionary histories of bacterial species.

However the knowledge of the rate and tract length of homologous bacterial recombination is still limited. Therefore a novel method to quantify homologous recombination by reconstruction of haplotypes in described in **Chapter 3**.

Chapter 3

Bacterial recombination analysis based on haplotype construction

3.1 Background

Recombination is a key mechanism that drives the architecture and evolution of bacterial genomes. Recombination in bacteria, in contrast to eukaryotes where recombination involves the process of reciprocal exchange of genetic material between homologous chromosomes, results in the addition of DNA, homologous or non-homologous, to another genome. Recombination in bacteria occurs through unidirectional transfer of genetic material from the donor to the recipient cell by three different mechanisms: transformation (novel genes appear in bacteria by taking up external DNA molecules from the environment), transduction (bacteria receive DNA from bacteriophage) and conjugation (the exchange of DNA is carried out by physical contact between two bacteria). Non-homologous recombination occurs when a foreign DNA segment is inserted into a position in the host genome. This event is called lateral (or horizontal) gene transfer [68-69].

Bacterial recombination does not occur in every generation and its frequency depends on the successful rate of DNA exchange as well as different biological and ecological factors[70]. Recombination events can be traced by either hybridization or sequencing technologies that are able to detect sequence variants or insertions of novel DNA. Prior to the advent of next generation sequencing methodology low resolution maps of recombination were constructed until the 1980s by genetic means and later by Sanger sequencing of a few housekeeping loci. The latter method called ‘Multilocus sequence typing’ (MLST) provided an overview of global recombination rates in a bacterial species. With more and more sequencing data sets at hand, it now becomes feasible to identify recombination rates and recombination breakpoints from high-throughput whole genome sequencing rather from multiple short regions.

3.2 About the manuscript

Here I introduce a novel method for the analysis of recombination breakpoints in bacterial genomes. Blocks of conserved sequence contigs are detected from pairwise comparison of bacterial genomes. The sequence of the core genome, i.e. the part of the genome that is conserved among all clone types of a bacterial species, is searched for single nucleotide sequence variants (SNPs) that are present in at least two genomes derived from isolates of spatiotemporally unrelated habitats. Pairwise alignment of genome sequences identifies the sequence of syntenic SNPs that are then converted from the genome coordinates into the physical length of nucleotide sequence. By performing pairwise comparisons of all genome sequences in the sample, the output is the distribution frequency of identical sequence length in the bacterial species of interest. In analogy to the terminology in eukaryotes, we simply define the 'sequence of syntenic SNPs' as a 'haplotype'. Phylogenetic trees are then constructed based on the criterion of the 'number of haplotypes shared between two bacterial strains'. This approach has been applied to recombination analysis of a) 20 *P. aeruginosa* genomes that are representative for the major 20 clone types in the population (see Chapter 4) and b) 100 genomes of isolates belonging to the two major *P. aeruginosa* clones named C and PA14 (see Chapter 5). Moreover, I analyzed genomes of the second most prominent pathogen in cystic fibrosis airways, *Staphylococcus aureus* (see below).

Author's contribution.

The author conceived the algorithm, wrote and set up the software pipeline, performed all subsequent genome analyses and drafted the manuscript. Genome sequences were either generated in-house or were downloaded from the NCBI web-site.

More information regarding the new algorithm can be found in ***Appendix2****.*

For Tables, please refer to the DVD attached to the thesis.

Bacterial recombination analysis based on haplotype construction.

Patricia Morán Losada[1,*],Philippe Chouvarine[1,2] and Burkhard Tümmler[1,2]

[1] Clinical Research Group, 'Molecular Pathology of Cystic Fibrosis and Pseudomonas Genomics', OE 6710, Hannover Medical School, Hannover D-30625, Germany.

[2] Biomedical Research in Endstage and Obstructive Lung Disease (BREATH), German Center for Lung Research, Hannover, Germany.

Email Patricia Morán Losada : MoranLosada.Patricia@mh-hannover.de

Email Philippe Chouvarine : Chouvarine.Philippe@mh-hannover.de

Email Burkhard Tümmler : Tuemmler.Burkhard@mh-hannover.de

* To whom correspondence should be addressed. Tel: +49 511 532-7838; Fax: +49 511 532-6723; Email: MoranLosada.Patricia@mh-hannover.de

Abstract

Background: The recent developments in next generation sequencing technologies enable new possibilities for the analysis of large volumes of DNA sequences. New statistical methods are required to detect signatures of natural selection in genomic data. Analyses of single nucleotide polymorphisms (SNPs) in the DNA of a species allows for the identification of 100% conserved stretches of sequences called haplotypes in diploid organisms which can provide information in addition to the established recombination methods. Using a matrix based binary algorithm we introduce a new approach for 'haplotype' analysis of the bacterial genome which is generally applicable to bacterial population genetics.

Results: Haplotypes defined by the length of syntenic segments with identical SNPs were derived from pairwise comparisons of bacterial genomes. Two matrices were constructed that contained columns of all quality-controlled SNPs ordered by genome position of the reference and rows of the bacterial isolates of interest respectively. SNP syntenies were extracted from pairwise comparisons of rows and converted into physical length. The outcome is the distribution of the length of haplotypes in the analyzed strain sample that can be exploited to visualize the relatedness of clades in a tree. This approach was applied to genome sequences of *Staphylococcus aureus* and *Pseudomonas aeruginosa* strains.

Conclusions: Pairwise genome comparisons of SNP synteny yield information about linkage and recombination in the core genome as a measure of the population structure of a taxon or a clonal complex.

Keywords: Bacterial recombination, haplotypes, *Staphylococcus aureus*, *Pseudomonas aeruginosa.*

Background

Recombination is a fundamental process in bacterial evolution that generates genome variation. Bacteria are prokaryote organisms that reproduce only clonally where DNA transfer is unidirectional and always independent of reproduction. However they occasionally exchange foreign fragments of DNA horizontally through one of three different processes: transduction (DNA introduced by bacteriophage), conjugation (DNA introduced by plasmids), transformation (uptake of free DNA from the environment) [1–4].

Based on the type of DNA which has been transferred, recombination can generate two outcomes, homologous and non-homologous recombination [5]. Homologous recombination occurs when the new variation is limited to a new allele, in other words, the DNA from the donor cell replaces its homologous allele in the recipient cell. This type of recombination requires incoming DNA to be highly similar to the recipient DNA. Non-homologous recombination happens when a novel gene or fragment of DNA is transferred from the donor cell into the genome of the recipient cell. Non-homologous recombination is often known as lateral gene transfer (LGT).

Bacterial genomes are composed of a core genome, which contains genes that are shared by all strains of the species, and an accessory genome, consisting of non-essential genes that might or might not be present in a given strain [6].

Multiple methods have been developed to identify genetic recombination in bacterial genomes. Most studies used multilocus sequence typing (MLST) [7, 8], which identifies all single nucleotide polymorphisms (SNPs) of seven housekeeping genes, giving a specific allelic profile to each sequence type. Alternatively, with next generation sequencing becoming a rapid and affordable technology, linkage and recombination can be deduced from whole genome SNP comparison. This approach is a timely topic in eukaryotic genetics where the genome-wide reconstruction of haplotypes provides comprehensive information about genome organization and diversity at the species and subspecies level [9–12].

Bacteria are haploid organisms, but in analogy to studies in diploid organisms linkage and recombination can be deduced from pairwise genome comparisons of strains that belong to the same species or infrataxonomic ranks thereof. Here we describe a straightforward approach to calculate the stretches of shared identical sequence (‘haplotype’) in taxa or clonal complexes from combinatorial pairwise whole genome SNP comparisons as a measure of the relatedness of strains or clonal complexes. One genome is taken as the reference and all SNPs seen in at least two strains are used for the analysis. When differentiated by genome position, gradients of recombination frequency within the core genome are ascertained. We illustrate the approach with a data set of *Staphylococcus aureus* and *Pseudomonas aeruginosa* genomes.

Methods

P. aeruginosa and *S.aureus* genome sequences

Representative strains of the 20 most common clonal complexes in the *P. aeruginosa* population were sequenced on an Illumina Genome Analyzer II [13]. Reads were aligned to the *P. aeruginosa* PAO1 reference genome (NC_002516.2) using the software SARUMAN (version 1.0.7) [14] with a maximum of 8 mismatches per read. SNPs were extracted with samtools and the core genome SNPs identified. Forty-one genome sequences of *S. aureus* were downloaded from the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) in March 2015. SNPs were extracted with the open source software package MUMmer, version 3.0 [15]. MUMmer's alignment and SNP calling utilities (nucmer and show-snp) were used with the default parameters. The genome of strain "Newman" [16] was taken as reference. Mutations occurring only in a single *P. aeruginosa* or *S. aureus* strain were filtered out.

Finally, 58 genome sequences of *P. aeruginosa* clone C isolates [17, 18] were sequenced on a SOLiD 5500XL system with 75 bp read length. Reads were aligned to PAO1 reference genome using the program NovoalignCS (www.novocraft.com). SNPs were extracted using samtools.

Haplotype definition

Different types of methods have been proposed for defining haplotypes. They can be classified into two groups: the first group uses the pairwise linkage disequilibrium to identify recombination areas [19] and the second group defines haplotypes as blocks with limited haplotype diversity [20]. In line with the latter definition, we define a haplotype as a block of syntenic SNPs that are shared by two genomes.

Haplotype analysis algorithm

To simplify the understanding for the reader we will explain the algorithm based on the 41 *S. aur*eus genomes. To identify haplotypes, a matrix was created where each column represents a SNP ordered by genome position and each row contains a strain (reference and 40 other strains). The value of 1 was assigned to the alternative nucleotide and the value of 0 was assigned to the nucleotide matching with the reference (Figure 1A). SNPs found in only one strain were excluded.

To perform all $\binom{41}{2}$ pairwise comparisons a second matrix was created whereby each row represents the comparison between two strains and each column represents a SNP position (Figure 1B). To identify haplotypes, the matches of identical values were counted in each row until the first mismatch (Figure 1B). All haplotypes of 820 pairwise comparisons were sorted by number to evaluate the frequency distribution. Alternatively, the physical length of each haplotype was calculated from the genome positions of the SNPs. Since the first and last positions of a *n* SNP haplotype defined by the SNP contig $SNP_m \ldots SNP_{m+n}$ are located within the

intervals SNP_{m-1}-SNP_m and SNP_{m+n}-SNP_{m+n+1}, respectively, the start and stop positions of a haplotype were arbitrarily assigned to the midpoints of the intervals.

To construct a phylogenetic tree, the haplotypes of all pairwise comparisons were sorted by decreasing number of successive SNPs and converted into ranks. Each pairwise genome comparison was then represented by the sum of rank numbers of their haplotypes which was used as input (Figure 1C) to the program SplitsTree (version 4.0) [21] to generate a neighbor joining tree (Figure 2).

Results and discussion

There were 136,258 SNPs and a total number of 8,704,567 haplotypes in the *S. aureus* genome data set, whereby the largest haplotype consisted of 2,450,522 nucleotides. In the *P. aeruginosa* genomes 113,172 SNPS and 3,779,224 haplotypes were identified, whereby the largest haplotype had 112,776 nucleotides between unrelated clones.

Recombination is studied in the Figure 3 which shows the frequency distribution of haplotypes as a function of number of consecutive syntenic SNPs in paired comparison of strains (N). The analysis shows that the median length of paired conserved sequences is about 5 syntenic SNPs for *P. aeruginosa* and *S. aureus.*

The analysis of *P. aeruginosa* clone C isolates and ST5 (ECT-R2, Mu3, Mu50 and N315) strains of *S. aureus* revealed a median length of paired conserved sequences of 10 and 33 syntenic SNPs, respectively.

Figure 4 shows the frequency of syntenic SNPs SNP_m…SNP_{m+n} as a function of the number *n* of successive SNPs. Most SNP contigs are shorter than 100 nucleotides indicating extensive sequence diversity at many loci in the two genomes of the taxa. Multiples of three nucleotides 3p are more frequent than 3p+1 and 3p+2 which reflect the overrepresentation of SNPs at the third codon position. The reader is reminded that haplotypes are longer than these SNP contigs because start and stop positions of the haplotype reside at genome positions somewhere between SNP_{m-1}-SNP_m and SNP_{m+n}-SNP_{m+n+1}, respectively. The median value for this extra sequence can be roughly estimated from the inverse proportion of the global sequence diversity in the core genome of individual strain pairs, i.e. 324 bp for *S. aureus* and 207 bp for *P. aeruginosa.*

We mapped all haplotypes onto the *S. aureus* and *P. aeruginosa* reference genomes (Figure 4). Short and long haplotypes were evenly distributed along the chromosome pairs in both taxa. The majority of haplotypes were shorter than 2,000 bp (99.4% in *S. aureus* and 99.3% in *P. aeruginosa)*. The median haplotype length was 51 bp for *S. aureus* and 99 bp for *P. aeruginosa* (Table 1). Very long haplotypes were only detected among *S. aureus* strain pairs, the largest one of 2,450,522 bp starting at position 37 of the reference genome (Figure 5A). The closer relatedness of subgroups of *S. aureus* strains was visualized in the neighbor joining tree (Figure 2A). Two pairs, three trios, one quartet, one septet and a further ten strains segregated into

clonal complexes. Consistent with this assignment, intraclonal pairs made up the 100 largest *S. aureus* haplotypes (Figure 6A). If we confined haplotype mapping to a single *S. aureus* clone, previously unnoticed gradients of haplotype length became apparent. As shown in Figure 7, the frequency of haplotypes was similar throughout the chromosomes of cluster 2 strains (Table2), but this continuous gradient was interrupted in three small regions characterized by very short haplotypes. These regions represent hot spots of mutation and/or recombination.

The *P. aeruginosa* strain panel was devoid of very long haplotypes (Figure 5B) consistent with the selection of the strains to represent the 20 most common clonal complexes of the *P. aeruginosa* population [13]. Based on haplotype relatedness, the strains segregated into one large cluster, one smaller cluster and one outlier (B420) (Figure 2B). The tree corresponds with that of whole genome comparisons based on single SNP diversity in the core genome [13]. *P. aeruginosa* strains 1BAE and 3C2A were the most closely related strains in the panel (Figure 2), and correspondingly 87 of the hundred largest haplotypes were assigned to this strain pair (Figure 6B). These two clones are more related to each other than the average randomly selected clone pair and probably emerged recently from a common ancestor.

Finally haplotypes given in physical length were determined for major clonal complexes. We chose isolates of the most common clone in the *P. aeruginosa* population, named clone C [18] and isolates of the pandemic healthcare-associated methicillin-resistant *S.aureus* (MRSA) clone ST5 (ECT-R2, Mu3, Mu50 and N315) (Figure 3B). The median haplotype lengths were 4.2 kbp for *S. aureus* ST5 and 99 kbp for *P. aeruginosa* clone C. The higher intraclonal relatedness of clone C compared to ST5 strains showed up by both the lower number of SNPs and the higher physical haplotype length. This data demonstrates that physical length rather than the number of syntenic SNPs provides a true estimate of the relatedness of strains within clones or species.

Conclusions

We have introduced a new algorithm for estimating intraspecies or intraclonal bacterial relatedness by genome-wide pairwise comparison of SNPs. In this approach, haplotypes are defined by the number of consecutive SNPs shared by two strains. This genetic entity is a measure for the size of linkage groups in the bacterial core genome. Related strains share a larger portion of longer haplotypes than unrelated strains. The distribution of haplotype length along the chromosome highlights the spatial distribution of recombination frequency. In our strain panels haplotype frequency was evenly distributed along the chromosome pairs suggesting unrestricted gene flow between clonal complexes by recombination.

The spatial resolution of haplotype analysis is determined by the number and map position of the SNPs. Start and stop positions of a haplotype are inherently unknown so that the physical length of an individual haplotype cannot be calculated with certainty. This error is of course most relevant for short haplotypes if the size of the syntenic SNP contig is smaller than or within the same range as the flanking intervals to the adjacent SNPs. However, in case of global genome comparisons this error becomes negligible if the start and stop positions of all haplotypes are

assigned to the midpoints of the intervals.

Availability

The developed algorithm for the identification and analysis of haplotypes is available for download from:
http://genomics1.mh-hannover.de/software/haplotypes_identification.pl.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PML developed the algorithm, wrote the script, and generated graphs using R packages. All authors contributed to writing the manuscript. PC and BT provided supervision and advice.

Acknowledgments

This work was partially funded by a grant from the Deutsche Forschungsgemeinschaft (SFB 900, Z1) to B.T. and supported by the Hannover Biomedical Research School (HBRS) and the Center for Infection Biology (ZIB). P.M.L. is a member of the ZIB.

References

1. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann M-L, Brüssow H: **Phage as agents of lateral gene transfer**. *Current opinion in microbiology* 2003, **6**:417–424.

2. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15**:589–594.

3. Ochman H, Lerat E, Daubin V: **Examining bacterial species under the specter of gene transfer and exchange.** *Proc Natl Acad Sci U S A* 2005, **102 Suppl 1**:6595–6599.

4. Smith JM, Smith NH, O'Rourke M, Spratt BG: **How clonal are bacteria?** *Proceedings of the National Academy of Sciences* 1993, **90**:4384–4388.

5. Vos M: **Why do bacteria engage in homologous recombination?** *Trends Microbiol* 2009, **17**:226–232.

6. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R: **Microbiology in the post-genomic era.** *Nat Rev Microbiol* 2008, **6**:419–430.

7. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **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.

8. Narra HP, Ochman H: **Of what use is sex to bacteria?** *Current Biology* 2006, **16**:R705–R710.

9. Genome of the Netherlands Consortium: **Whole-genome sequence variation, population structure and demographic history of the Dutch population.** *Nat Genet* 2014, **46**:818–825.

10. Gasbarra D, Kulathinal S, Pirinen M, Sillanpää MJ: **Estimating haplotype frequencies by combining data from large DNA pools with database information.** *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8**:36–44.

11. Kirkpatrick B, Armendariz CS, Karp RM, Halperin E: **HAPLOPOOL: improving haplotype frequency estimation through DNA pools and phylogenetic modeling.** *Bioinformatics* 2007, **23**:3048–3055.

12. Patterson M, Marschall T, Pisanti N, Iersel L van, Stougie L, Klau GW, Schönhuth A: **WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads.** *J Comput Biol* 2015, **22**:498–509.

13. Hilker R, Munder A, Klockgether J, Losada PM, Chouvarine P, Cramer N, Davenport CF, Dethlefsen S, Fischer S, Peng H, Schönfelder T, Türk O, Wiehlmann L, Wölbeling F, Gulbins E, Goesmann A, Tümmler B: **Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment.** *Environ Microbiol* 2015, **17**:29–46.

14. Blom J, Jakobi T, Doppmeier D, Jaenicke S, Kalinowski J, Stoye J, Goesmann A: **Exact and complete short-read alignment to microbial genomes using Graphics Processing Unit programming.** *Bioinformatics* 2011, **27**:1351–1358.

15. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**:R12.

16. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K: **Genome sequence of Staphylococcus aureus strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands.** *J Bacteriol* 2008, **190**:300–310.

17. Cramer N, Klockgether J, Wrasman K, Schmidt M, Davenport CF, Tümmler B: **Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs.** *Environ Microbiol* 2011, **13**:1690–1704.

18. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Köhler T, Delden C van, Weinel C, Slickers P, Tümmler B: **Population structure of Pseudomonas aeruginosa.** *Proc Natl Acad Sci U S A* 2007, **104**:8101–8106.

19. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225–2229.

20. Zhang K, Deng M, Chen T, Waterman MS, Sun F: **A dynamic programming algorithm for haplotype block partitioning.** *Proc Natl Acad Sci U S A* 2002, **99**:7335–7339.

21. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23**:254–267.

Table and figure legends

Table 1. Distribution of haplotypes based on their physical length.

Table 2. **Top largest haplotype blocks in *Staphylococcus aureus* and their classification in the different clusters.** Most of the largest haplotype blocks belong to the main clusters (cluster1 and 2).

Figure 1. Schematic representation of haplotype identification and phylogenetic tree generation. (A) The matrix shows the presence or absence of all SNP positions in all strains. (B) Example of haplotype (grey) identification (based on the first matrix), counting the successive matches of identical values for each pair until the first mismatch. (C) Distance matrix to generate the phylogenetic tree. Each value represents the sum of all rank numbers based on the number of syntenic SNPs for each haplotype.

Figure 2. **Phylogenetic tree of *S. aureus* (A) and *P. aeruginosa* (B), based on neighbor-joining matrix distance.**

Figure3. Distribution haplotypes based on syntenic SNPs length. (A) It shows the distribution of haplotypes in clonally unrelated *S. aureus* and *P. aeruginosa* strains

(B) It shows the distribution of haplotypes in Clone C and ST5 clonal isolates.

Figure 4. Distribution haplotype blocks based on the nucleotides length. (A) It shows the distribution of haplotyes in *Staphylococcus aureus* where the majority of haplotypes have a log(length) less than 3. Two main parallel curves define the distribution of haplotype lengths.

(B) The graph shows the distribution of haplotypes in *Pseudomonas aeruginosa.*

Figure 5. Distribution of the physical length of (A) *S. aureus* and (B) *P. aeruginosa* haplotypes along the genome.

Figure 6. Localization of the 100 largest haplotypes in genome segments of 0.5 Mbp [*S. aureus* (A)] or 1 MBp [*P. aeruginosa*] .

Figure 7. Distribution of the physical length of *S. aureus* cluster 2 haplotypes along the genome.

Figure 1

A

Genome_position	SNP_{x1}	SNP_{x2}	SNP_{x3}	SNP_{x4}	SNP_{x5}	SNP_{x6}	…	SNP_{xn}
StrainA	0	0	1	0	0	1	…	1
StrainB	0	0	1	1	0	1	…	0
StrainC	0	0	1	0	0	1	…	1
StrainD	0	0	1	1	1	0	…	0
StrainE	1	1	1	1	1	1	…	1
…								
Reference	0	0	0	0	0	0	…	0

B

Genome_position	SNP_{x1}	SNP_{x2}	SNP_{x3}	SNP_{x4}	SNP_{x5}	SNP_{x6}	SNP_{xn}
StrainA/B	3	2	1	0	2	1	...
StrainA/C	6	5	4	3	2	1	...
StrainA/D	3	2	1	0	0	0	...
StrainA/E	0	0	1	0	0	1	...
StrainA/Ref	2	1	0	2	1	0	...
StrainB/C	3	2	1	0	2	1	...
...							

C

	StrainA	StrainB	StrainC	StrainD	StrainE	StrainF
StrainA	0	-	-	-	-	-
StrainB	AB	0	-	-	-	-
StrainC	AC	BC	0	-	-	-
StrainD	AD	BD	CD	0	-	-
StrainE	AE	BE	CE	DE	0	-
StrainF	AF	BF	CF	DF	EF	0

Figure 2

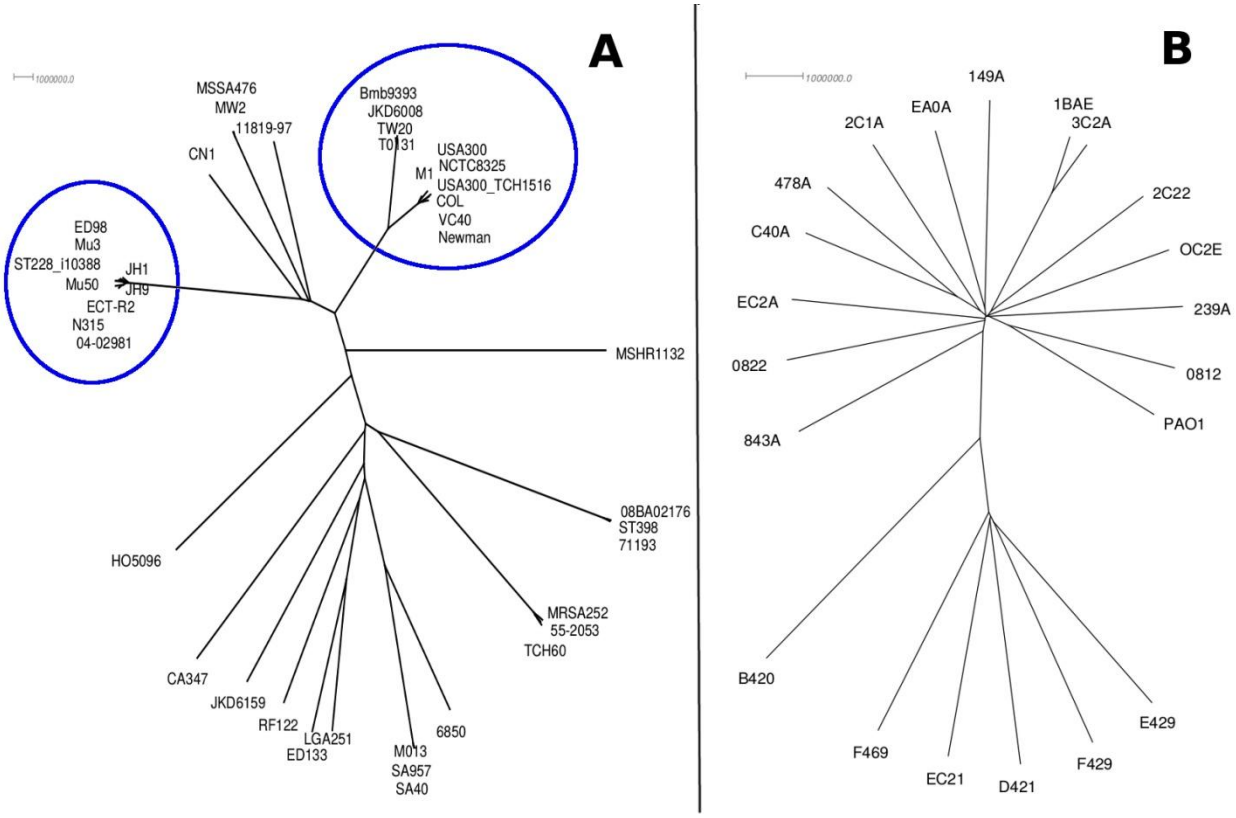


Figure 3

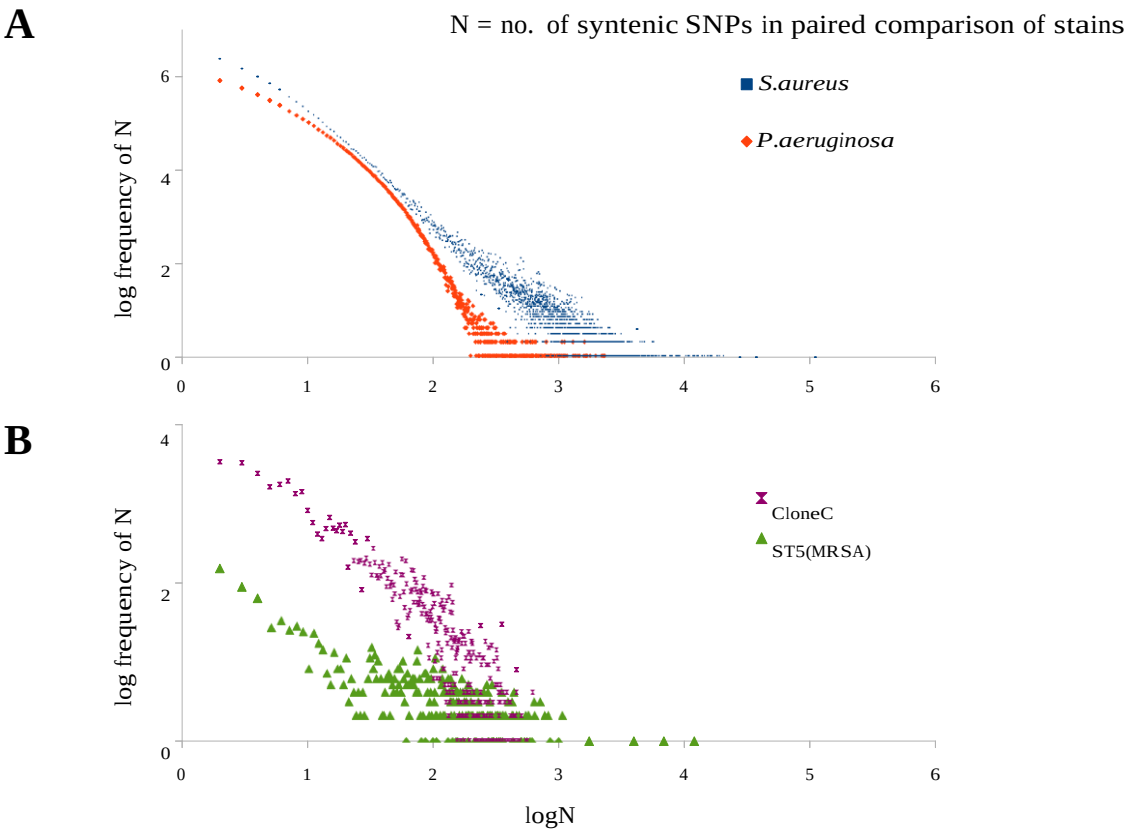


Figure 4

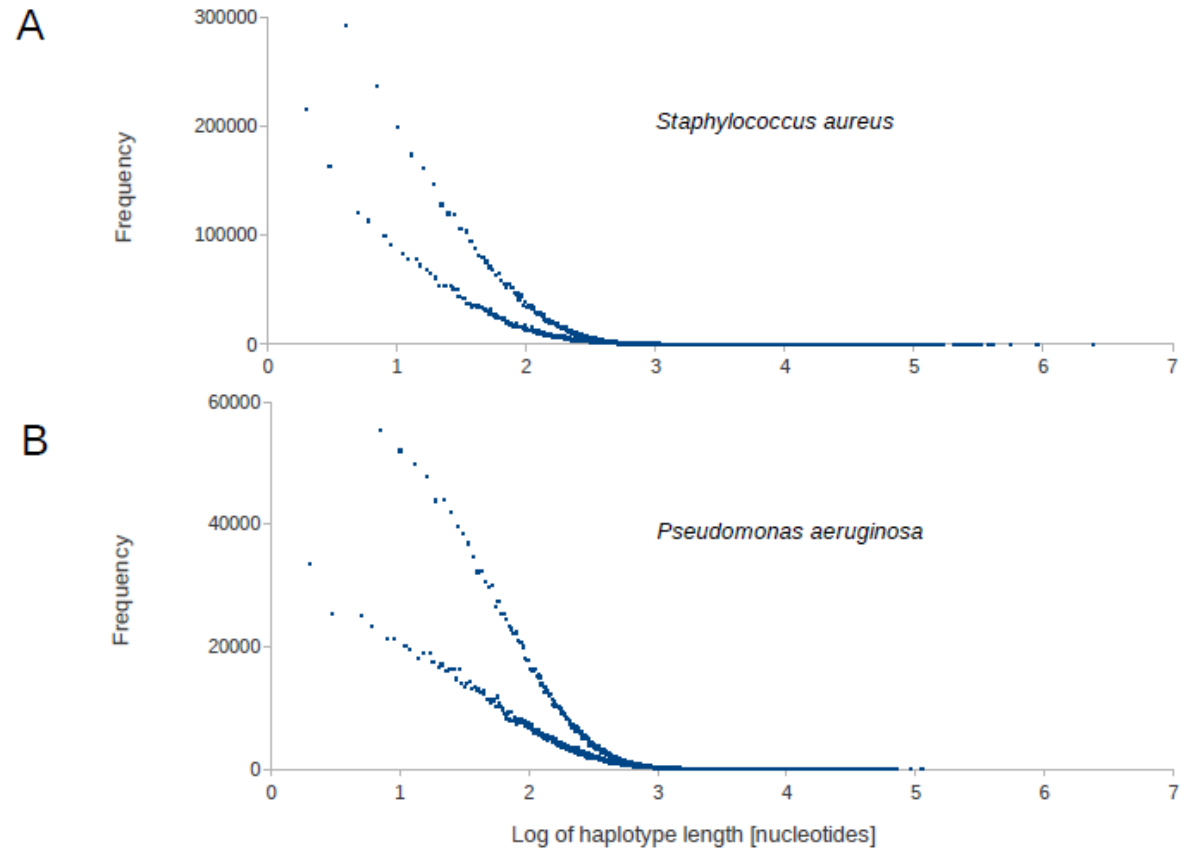


Figure 5

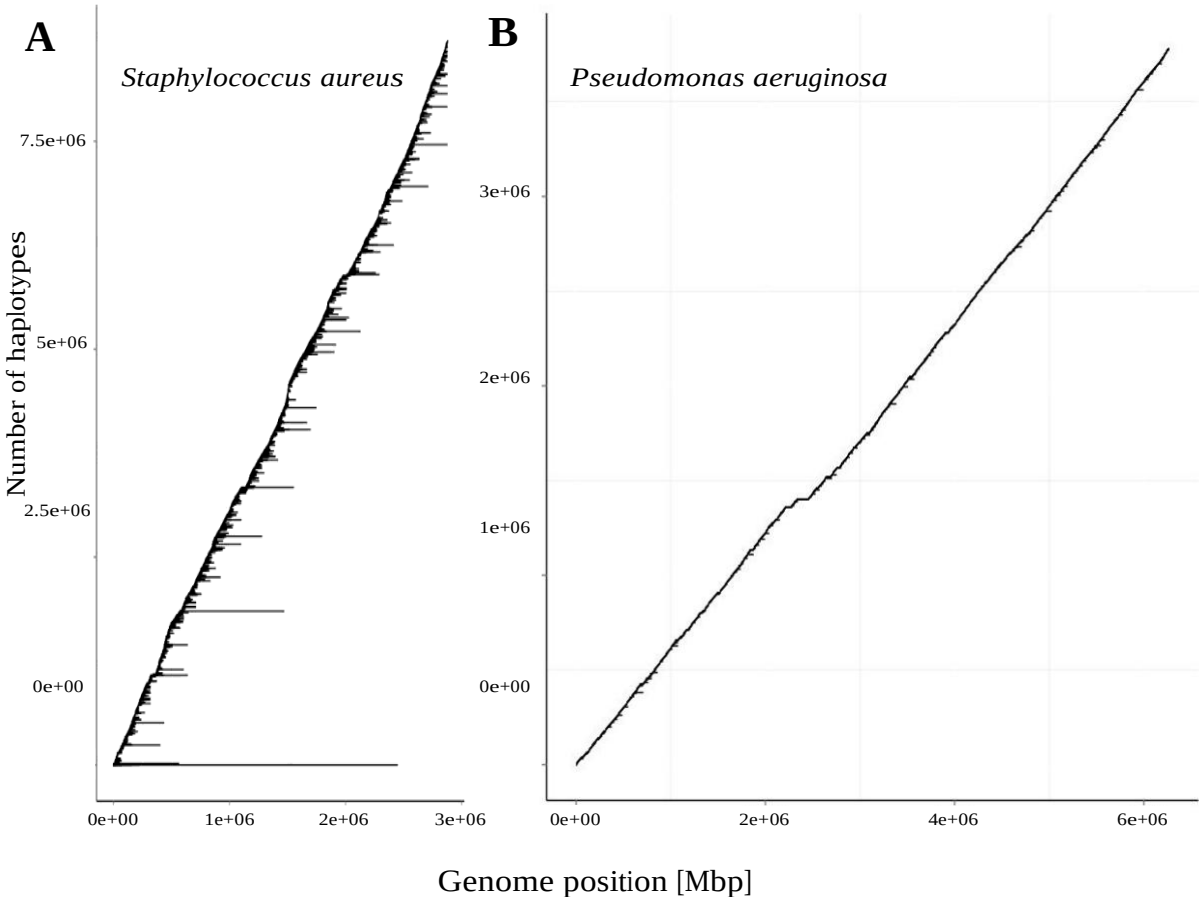


Figure 6

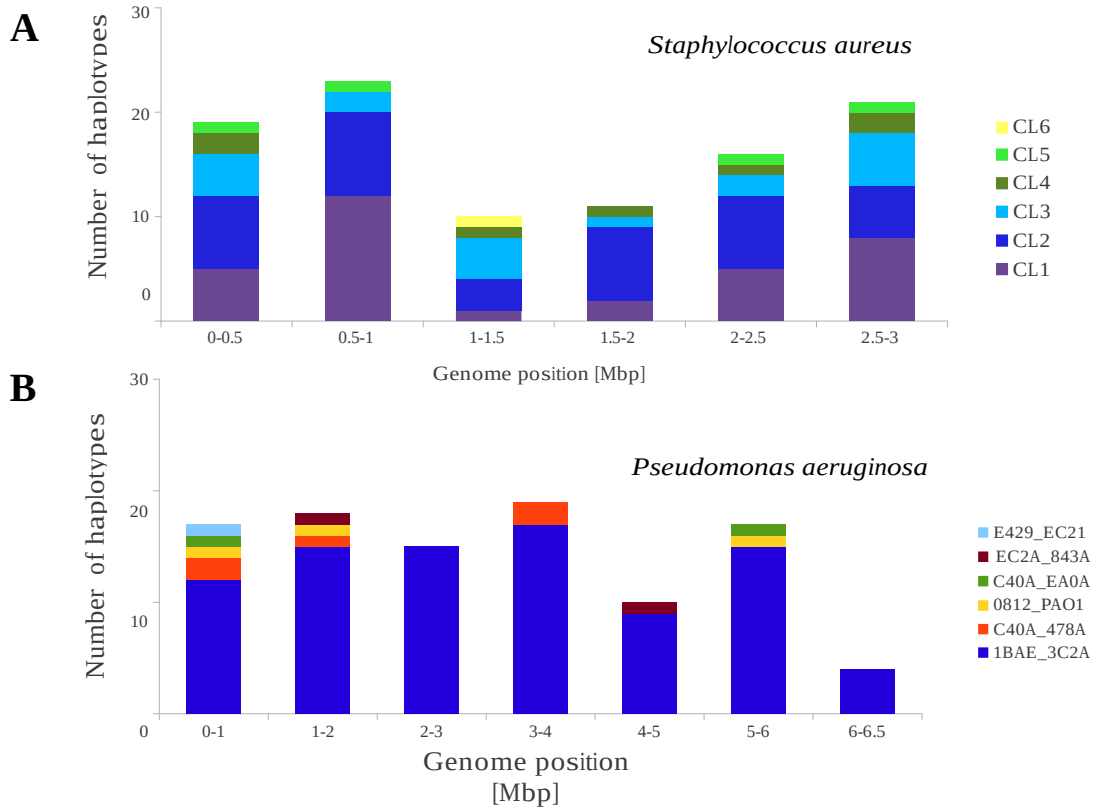
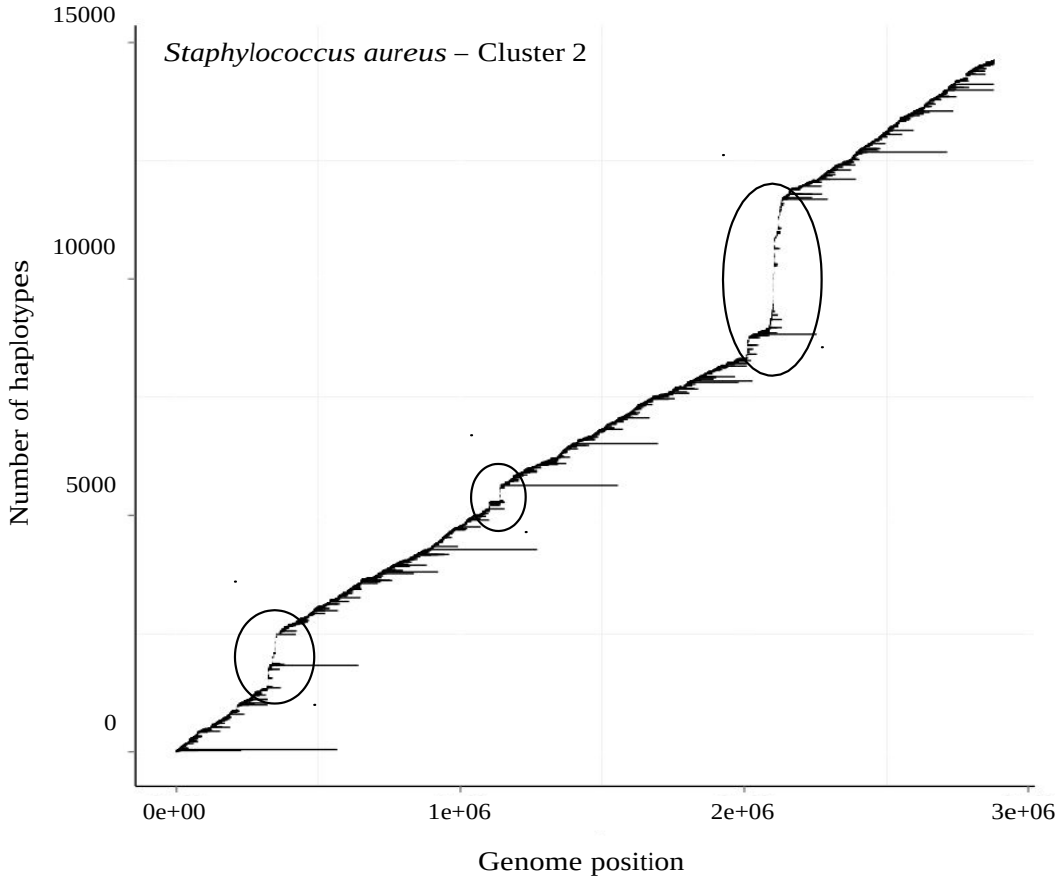


Figure 7


Staphylococcus aureus – Cluster 2
15000
10000
5000
0
Number of haplotypes
0e+00
1e+06
2e+06
3e+06
Genome position

Chapter 4

Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome

4.1 Background

Studies of population genomics contribute information on the population structure and its genetic variation. Quantifying the rate of mutation and recombination is essential for understanding epidemiological changes occurring in a population.

Studying bacterial populations also allows us to characterize the strains of pathogenic species; but it is also important to understand the genetic relationships existing between pathogenic strains (causing disease), and non-pathogenic strains of the same species[71-72]. The comparison between both populations may help us to explain the origins of pathogenic strains and identify genetic differences between them.

If there were no chromosomal recombination in bacteria, their populations would have a clonal structure whereby genetic variation would be given by mutation[73]. However, in the real world recombination occurs in bacteria and promotes adaptive evolution providing certain allele combinations that can be exchanged between different bacterial species, such as antibiotic resistance genes, virulence determinants, nitrogen fixation, and so on[74-76].

Pseudomonas aeruginosa is an opportunistic Gram-negative pathogen that colonizes a wide range of niches, it is able to survive in aquatic, animal and human-associated habitats[77]. As an opportunistic pathogen in humans, *P. aeruginosa* has become one of the main causes of infections in patients with advanced stages of chronic obstructive pulmonary disease and individuals suffering from cystic fibrosis. *P.aeruginosa* adaptability and metabolic versatility is the fundamental key to its survival (it can grow under aerobic and anaerobic conditions, use multiple carbon sources for energy, form biofilms and resist many antibiotics)[78-80].

Next generation sequencing has made comparative genomic analysis of *P. aeruginosa* possible as well as a better understanding of its recombination mechanism.

4.2 About the manuscript

This research project has the aim of resolving the pangenome of *P. aeruginosa* and identifying the genomic variation between strains from environmental and disease habitats. Sequence analyses were performed on the 15 most frequent clonal complexes in the *P. aeruginosa* population and 5 most common environmental clones.

To investigate the role of recombination in genome mobility of *P.aeruginosa*, the novel method for the analysis of recombination breakpoints in bacterial genomes described in Chapter 3 was applied. All strain genomes were compared against the core genome of the reference PAO1 in pairwise comparisons. Only SNPs present in at least two of the 20 sequenced genomes were taken for the analysis avoiding *de novo* mutations present in a single strain.

The frequency distribution of the number of syntenic SNPs was determined and transformed into the physical length of sequence. 192,443 SNPs were identified in the 210 paired comparisons suggesting that the median length of paired conserved sequence is 207 base pairs. We found two clones more related to each other than the average, the clones 1BAE and 3C2A, which points to a recent origin from a common ancestor.

This analysis revealed that the pangenome consists of a conserved “core” genome of about 4,000 genes shared among all members of the study, “accessory” genomic elements of a further 10,000 genes that are present in some but absent in other strains of *P. aeruginosa* and around 30,000 rare genes present in only few strains.

Author’s contribution.

The paper was conducted by Rolf Hilker, Antje Munder and Jens Klockgether.
I performed the recombination analysis of the P. aeruginosa pangenome as well as the graphs and tables describing the recombination process.
For Tables and Supplementary material, please refer to the DVD attached to the thesis.

environmental microbiology

Environmental Microbiology (2015) **17**(1), 29–46 doi:10.1111/1462-2920.12606

Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment

Rolf Hilker,[1†] Antje Munder,[2†] Jens Klockgether,[2†] Patricia Moran Losada,[2] Philippe Chouvarine,[2] Nina Cramer,[2] Colin F. Davenport,[2] Sarah Dethlefsen,[2] Sebastian Fischer,[2] Huiming Peng,[3] Torben Schönfelder,[2] Oliver Türk,[2] Lutz Wiehlmann,[2] Florian Wölbeling,[2] Erich Gulbins,[3] Alexander Goesmann[1] and Burkhard Tümmler[2,4*]

[1]*Department of Bioinformatics and Systems Biology, University of Giessen, Gießen D-35392, Germany.*

[2]*Clinical Research Group, 'Molecular Pathology of Cystic Fibrosis and Pseudomonas Genomics', OE 6710, Hannover Medical School, Hannover D-30625, Germany.*

[3]*Department of Molecular Biology, University Hospital Essen, University of Duisburg-Essen, Essen D-45122, Germany.*

[4]*Biomedical Research in Endstage and Obstructive Lung Disease (BREATH), German Center for Lung Research, Hannover, Germany.*

Summary

The population genomics of *Pseudomonas aeruginosa* was analysed by genome sequencing of representative strains of the 15 most frequent clonal complexes in the *P. aeruginosa* population and of the five most common clones from the environment of which so far no isolate from a human infection has been detected. Gene annotation identified 5892–7187 open reading frame (ORFs; median 6381 ORFs) in the 20 6.4–7.4 Mbp large genomes. The *P. aeruginosa* pangenome consists of a conserved core of at least 4000 genes, a combinatorial accessory genome of a further 10 000 genes and 30 000 or more rare genes that are present in only a few strains or clonal complexes. Whole genome comparisons of single nucleotide polymorphism synteny indicated unrestricted gene flow between clonal complexes by recombination. Using standardized acute lettuce, *Galleria mellonella* and murine airway infection models the full spectrum of possible host responses to *P. aeruginosa* was observed with the 20 strains ranging from unimpaired health following infection to 100% lethality. Genome comparisons indicate that the differential genetic repertoire of clones maintains a habitat-independent gradient of virulence in the *P. aeruginosa* population.

Introduction

Pseudomonas aeruginosa is a ubiquitous metabolically versatile gammaproteobacterium that can thrive at low densities within the range of 4–42°C in inanimate aquatic habitats and can colonize the surface of animate hosts ranging from worms and flies to plants and mammals (Ramos, 2004–2010a,b). Being an opportunistic pathogen, *P. aeruginosa* causes a wide range of syndromes in humans that can vary from local to systemic, subacute to chronic, and superficial and self-limiting to life-threatening.

The *P. aeruginosa* population has an epidemic structure (Pirnay *et al.*, 2009; Selezska *et al.*, 2012). By genotyping a large collection of strains from environmental and disease habitats with a custom-made multi-marker array, we have identified several hundred different clonal complexes, the majority of which are rare (Wiehlmann *et al.*, 2007; Cramer *et al.*, 2012). The 15 most frequent clones make up about 40% of the contemporary population. Members of the two major clones C (Römling *et al.*, 2005) and PA14 (Rahme *et al.*, 1995) were sampled from salt and fresh water, secluded national reserves, anthropogenically polluted sites, plants, wild and domestic animals, and acute and chronic human infections. In other words, the two global clones are everywhere. However, the next frequent clones preponderate for geographic areas and/or habitats. Numerous clones have no representative as yet among the subset of human infections, and conversely, clones that had caused outbreaks of nosocomial infection still lack an environmental isolate in our strain collection. These data suggest that the *P. aeruginosa* population consists of global and local generalists on one hand and niche specialists on the other. There even may exist an interclonal gradient of pathogenicity ranging from innocuous to highly virulent clones.

Received 14 March, 2014; accepted 5 July, 2014. *For correspondence. E-mail tuemmler.burkhard@mh-hannover.de; Tel. +49 511 5322920; Fax +49 511 5326723. †These authors contributed equally to this work.

In this study, we have analysed the genomes of 15 *P. aeruginosa* strains that are representative for the 15 most frequent clonal complexes in the *P. aeruginosa* population. Based on the hypothesis that clonal complexes may occur in the environment that cannot cause disease in humans, we added five more strains from soil, plants and aquatic habitats to this panel for genome sequencing, each of which representing a common clonal complex of which so far no isolate from a human niche has been detected. To address this issue of whether there is an association between clonal frame and virulence, we chose acute lettuce (Starkey and Rahme, 2009), *Galleria mellonella* (Pustelny *et al*., 2013) and murine airway infection models (Wölbeling *et al*., 2011) to compare the pathogenic capacity of the 20 strains. Under the strictly standardized conditions of these infection models, an unexpectedly large variability of virulence was discovered among the 20 strains.

The genome sequences of numerous *P. aeruginosa* strains have meanwhile been deposited in databases (Winsor *et al*., 2011). However, all sequenced strains apart from PA14 belong to uncommon clonal complexes in the bacterial population. Sequencing of our strain panel thus provided the opportunity to compare the gene content and sequence diversity among the most common clonal complexes and to estimate the gene pool of the pangenome of *P. aeruginosa*.

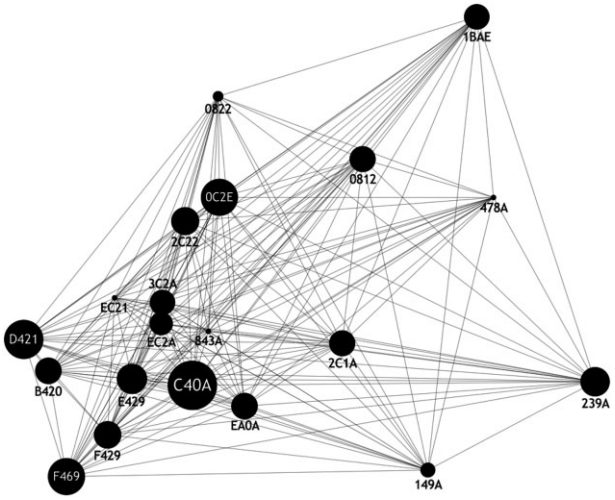


Fig. 1. Relatedness of the 20 sequenced strains. Starting from the GeneChip multimarker genotypes (Wiehlmann *et al*., 2007), the relatedness of strains was calculated in a 13-dimensional Manhattan space. The genetic distance was visualized by applying an optimized spring model for minimal stress in the GRAPHVIZ package (Gansner and North, 2000; Gansner *et al*., 2005). The localization of the 20 clonal lineages was based on the whole dataset of 1400 independent *P. aeruginosa* isolates. The size of the symbol represents the square root of the abundance of the clonal lineage in the *P. aeruginosa* population (Cramer *et al*., 2012).

Results and discussion

The P. aeruginosa *pangenome*

The 15 most common clonal complexes in the *P. aeruginosa* population were represented by isolates from the environment (2×), acute eye infection (1×), community-acquired pneumonia (1×), intubated patients (3×) and chronic airway infections of individuals with cystic fibrosis (CF) (5×) or chronic obstructive pulmonary disease (COPD) (3×). Five further environmental isolates were recovered from plants (2×), soil (1×), fresh (1×) and salt water (1×) respectively (Fig. 1; Table 1).

The median genome size of the 20 strains was determined to be 6.8 Mbp (range 6.4–7.4 Mbp; Table 1), which is larger than that of the completely sequenced *P. aeruginosa* genomes deposited in the Pseudomonas Genome Database (range 6.2–6.8 Mbp, median 6.4 Mbp) (Winsor *et al*., 2011). Annotation identified 5892–7187 open reading frames (ORFs; median 6381 ORFs) in the 20 individual genome sequences (Table 1). The 20 strains and reference strain PAO1 (Stover *et al*., 2000) share 4748 ORFs (Fig. 2), suggesting that the core genome common to all *P. aeruginosa* should consist of more than 4000 ORFs. Altogether 13 527 different ORFs were annotated in the 21 genomes (Table S1). As highlighted *pars pro toto* for four genomes in Fig. 3, the gene content is combinatorial in *P. aeruginosa* (Lee *et al*., 2006) which primarily reflects the combinatorial composition of the accessory genome with genomic islands (GIs) and insertions in regions of genomic plasticity (RGPs) (Mathee *et al*., 2008; Klockgether *et al*., 2011). Heatmaps visualize the variable presence and variable conservation of GIs and RGPs (Fig. 4A and B) in the 20 clonal complexes. None of 20 known GIs and only 16 of 245 known RGPs were present as complete copies in all 20 sequenced genomes.

Annotation identified 29–802 (median 167) genes that were detected in only one of the 20 genomes (Table S1). The major portion of these genes is organized in operons or DNA blocks of up to 312 ORFs. Most genes were assigned to the categories of nucleic acid metabolism, mobile genetic elements or hypotheticals. The disease isolates additionally carried a few paralogues of housekeeping enzymes and/or elements of mobility or secretion, the most spectacular example being an additional set of pilus biogenesis genes in the intubated patient isolate E429 (Table S1). The environmental isolates harboured larger sets of strain-specific genes than the clinical isolates, and these genes conferred numerous extra potential for bioenergetics, metabolism, transport and immunity such as a CRISPR-Cas system (Bondy-Denomy *et al*., 2013). Recurrent features were operons for the biogenesis of heme proteins and for the

Table 1. Genome characteristics of the sequenced *P. aeruginosa* strains.

Clone	Isolate	Source	Genome size (bp)	Number of predicted ORFs	Total number of variations	Substitutions	Insertions	Deletions	Intergenic SNPs	Synonymous SNPs	Non-synonymous SNPs	Insertions in genes	Deletions in genes
C40A	NN2	CF lungs, Germany	6 902 963	6583	24 861	24 509	147	205	13 961	8 545	2 288	25	42
D421	RN3	CF lungs, Germany	6 902 893	6412	55 876	55 192	331	353	30 599	20 104	5 052	48	73
F469	60P57PA	COPD airways, USA	6 910 555	6566	53 473	52 886	263	324	29 448	19 147	4 796	30	52
0C2E	RP1	CF lungs, Germany	6 914 674	6529	27 662	27 250	183	229	15 210	9 664	2 711	37	40
E429	15108/-1	Intubated patient, Spain	7 039 190	6585	55 148	54 494	303	351	30 118	19 918	5 001	50	61
239A	13121/-1	Intubated patient, France	6 915 596	6575	27 883	27 525	185	173	15 312	9 806	2 692	43	30
2C22	57P31PA	COPD airways, USA	6 519 939	6135	27 538	27 204	148	186	15 412	9 386	2 680	26	34
F429	A5803	Pneumonia, Germany	6 801 202	6397	55 306	54 627	331	348	30 229	19 953	5 015	43	66
B420	120SD3	River, Germany	6 368 472	5892	114 284	113 091	518	675	61 501	42 516	10 051	83	133
EA0A	39177	Keratitis, UK	6 629 320	6213	27 376	26 994	179	203	14 912	9 665	2 722	35	42
0812	27103	Intubated patient, France	6 652 994	6157	23 911	23 582	167	162	13 071	8 385	2 398	30	27
2C1A	18P17PA	COPD airways, USA	6 425 255	5959	28 342	27 938	174	230	15 546	9 922	2 808	31	35
1BAE	KK1	CF lungs, Germany	6 798 656	6381	26 548	26 286	100	162	14 595	9 301	2 612	15	25
3C2A	TR1	CF lungs, Germany	6 774 714	6386	26 639	26 371	107	161	14 726	9 273	2 598	16	26
EC2A	PT22	River, Germany	7 382 875	7187	27 045	26 762	101	182	14 590	9 730	2 683	17	25
EC21	100	Pacific Ocean, Japan	6 652 626	6285	52 989	52 521	210	258	28 975	19 132	4 806	30	46
843A	BP35	Pepper plant, India	6 351 657	5931	29 877	29 560	123	194	16 049	10 839	2 930	23	36
0822	E501	Tomato plant, Italy	7 037 971	6715	27 587	27 312	110	165	15 185	9 594	2 758	19	31
149A	120SB2	River, Germany	7 007 315	6536	28 458	28 156	130	172	15 494	10 143	2 772	18	31
478A	M41A.1	Soil, Colombia	6 370 109	6009	27 266	26 986	118	162	15 312	9 384	2 519	18	33

The first 15 strains represent the 15 most common clonal complexes in the *P. aeruginosa* population (sorted according to decreasing frequency). The lower five strains were from the most common clonal lineages without any clinical representatives found so far.

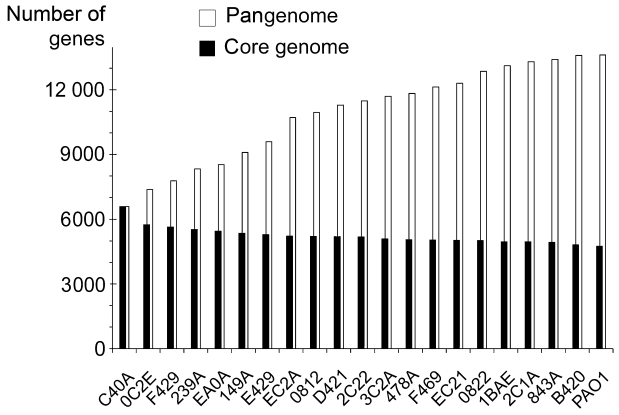


Fig. 2. The pangenome of the 20 sequenced strains representing the most common clonal complexes in the *P. aeruginosa* population. *P. aeruginosa* PAO1 was included as the reference strain. The coding genetic repertoire of the core genome and of the pangenome was constructed as follows: starting from the most frequent clone C (hexadecimal code C40A), core and pangenome were stepwise constructed by the addition of ORFs not present in the predecessor (pangenome) or by the substraction of ORFs absent in the successor (core genome). The numbers were obtained by the core and pangenome functions of EDGAR.

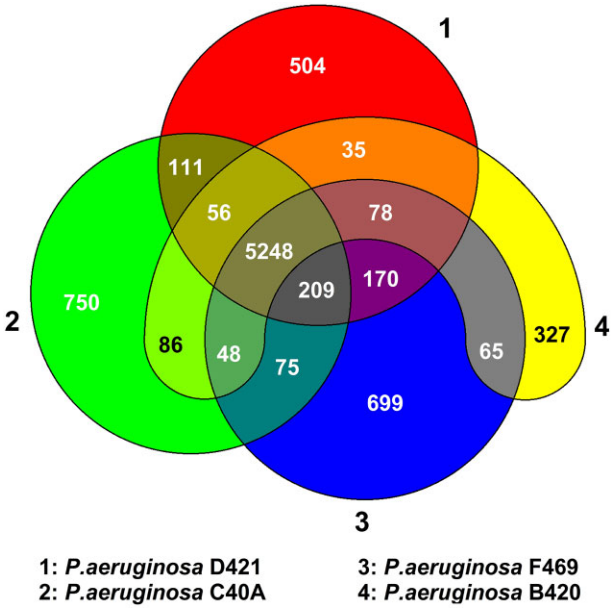


Fig. 3. Venn diagram of the number of strain-specific and shared ORFs in a panel of four *P. aeruginosa* strains consisting of the apathogenic B420 strain, the highly virulent F469 strain and the two strains belonging to the major clones C (C40A) and PA14 (D421).

transport and metabolism of amino acids and sulphur compounds.

Our genome datasets of the most frequent clonal complexes in the population suggest that the *P. aeruginosa* pangenome consists of a conserved core genome of at least 4000 genes, an accessory genome of common GIs and RGPs of about a further 10 000 genes and rare genes that are only present in few strains or clonal complexes. In this and other genome sequencing projects, dozens to hundreds of genes previously unknown in *P. aeruginosa* have regularly been observed whenever a strain of a yet uncharacterized clonal lineage was subjected to genome sequencing. Because more than 300 clonal complexes have been identified for *P. aeruginosa* to date, we can estimate a pool of at least 30 000 'private' genes that are rare or very rare in the *P. aeruginosa* population. Our empirical data fit perfectly with Koonin and Wolf's (2012) concept that a prokaryotic pangenome is made up of a small, highly conserved core, a much larger 'shell' of genes with limited conservation and a vast 'cloud' of rare poorly conserved genes.

Sequence diversity in P. aeruginosa

Mapping of the 20 genome sequences onto the *P. aeruginosa* PAO1 reference genome identified sequence variation between PAO1 and the respective strain at 23 911–114 284 positions of the PAO1 genome, which corresponds to a sequence diversity at the single nucleotide level of 0.38–1.82% (Table 1, Tables S2A and S2B). A portion of 8.8–10.1% of single nucleotide polymorphisms (SNPs) caused an amino acid substitution in ORFs. Based on the criterion of single nucleotide substitution (SNP) diversity, the clonal complexes were grouped into two clusters and one outlier (B420) (Fig. 5). The larger cluster including the most abundant clone C

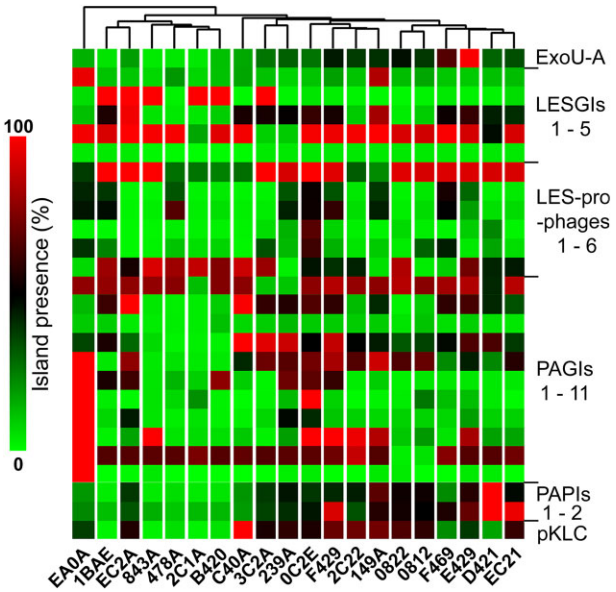


Fig. 4. Presence of known genomic islands and RGPs of the 20 sequenced genomes. Strains are grouped by similarity of their repertoire of genomic islands and RGPs. Coverage of an island or RGP is depicted in colour code ranging from lime (100% absent) via dark-green, black, crimson to light red (100% present).

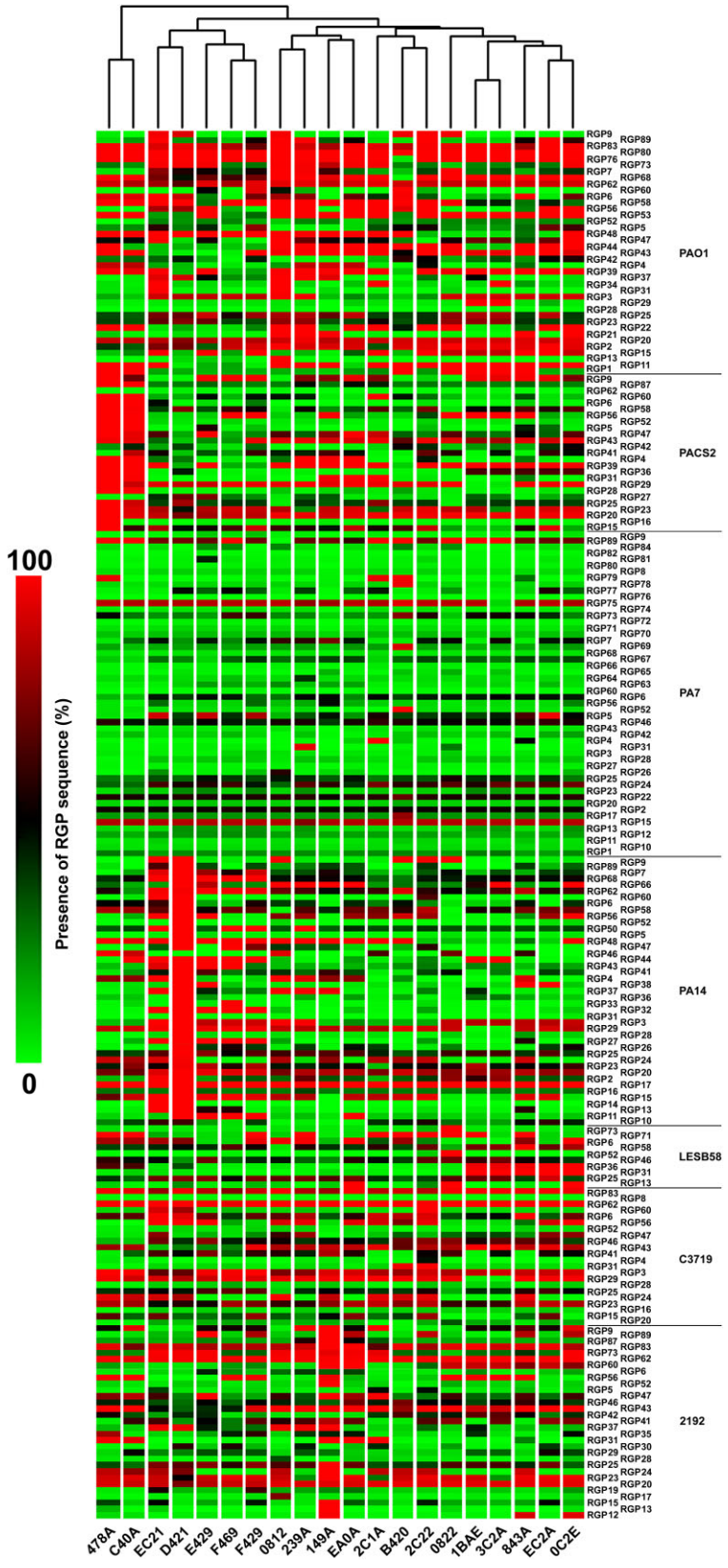


Fig. 4. *cont.*

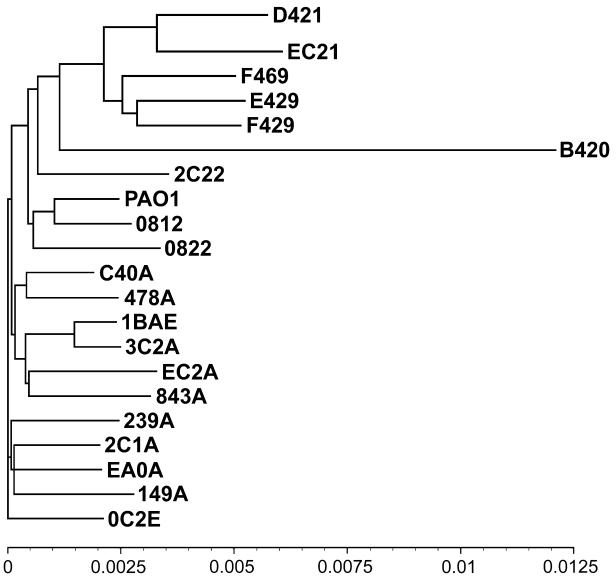


Fig. 5. Phylogenetic tree of the sequenced *P. aeruginosa* strains based on the paired comparison of SNPs in the core genome. The scale indicates the sequence diversity.

(C40A) differs at 0.38–0.48% positions from the PAO1 genome, whereas the smaller cluster including the second most abundant clone PA14 (D421) exhibits a sequence diversity of 0.85–0.88% (Table 1).

Most oligonucleotide insertions or deletions were in frame and caused the incorporation or loss of a single codon (Tables S2A and S2B). Single nucleotide out-of-frame mutations predominantly affected conserved hypotheticals. Deleterious frameshift mutations in functionally characterized genes were only detected in the 13 clinical isolates of our panel but in none of the seven environmental isolates. Recurrent loss-of-function hits were observed in gene clusters encoding the biosynthesis or regulation of flagella, pili, quinolones, the O-antigen of lipopolysaccharides, effectors of type III secretion, siderophores and their receptors and the biosynthesis of the antimetabolite L-2-amino-4-methoxy-trans-3-butenoic acid (Lee *et al*., 2010) (Tables S2A and S2B). This spectrum of mutations is consistent with a conversion of bacterial phenotype that often occurs during human infections, i.e. loss of motility, LPS deficiency and modulation of virulence, signalling and iron homeostasis (Döring *et al*., 2011).

Genome mobility

Lateral gene transfer and recombination shape the dynamics of bacterial genomes. The combinatorial repertoire of GIs and RGPs (Fig. 4) of the 20 strains indicates a continuous horizontal gene flow between the clonal complexes of *P. aeruginosa*. Numerous but not all strains, for example, shared complete copies of the LESprophage 1 and LESGI-4 (Winstanley *et al*., 2009) and harboured similar but not identical members of the PAGI-2 and pKLC102 island families (Klockgether *et al*., 2007).

To investigate the role of recombination in genome mobility of *P. aeruginosa*, all positions in the PAO1 core genome were marked wherein at least two of the 20 sequenced genomes a nucleotide substitution had been reliably identified. Then the frequency distribution of N, i.e. the number of syntenic SNPs, was determined along the core genome in paired comparisons of strains (Fig. 6) and converted into the physical length of sequence. The analysis of 192 443 SNPs in the 210 paired comparisons (Fig. S1) revealed that the median length of paired conserved sequence is 207 base pairs. Of the 110 blocks that are longer than 20 kbp in size, 78 were assigned to the pair of 1BAE/3C2A including the longest stretch of 90 949 bp (Table S3). These two clones are more related to each other than the average randomly selected clone pair and probably emerged recently from a common ancestor.

These whole genome comparisons of SNP synteny demonstrate that 75% of the blocks of identical sequence in the core genome that are shared by two clonal complexes are smaller than 350 base pairs. With the possible exception of two regions around 1.0 and 3.5 Mbp in the PAO1 genome (Fig. 6, Table S3), short blocks were evenly distributed along the chromosome suggesting unrestricted gene flow between clonal complexes by recombination. This conclusion is supported by the matching topology of trees irrespectively of whether the construction was based on syntenic SNPs – commonly termed a haplotype in case of diploid genomes – or on SNPs treated as independent singletons (Fig. 5).

Interclonal gradient of virulence in a murine airway infection model

Because the lower airways are the habitat where *P. aeruginosa* causes most morbidity and mortality in humans (Döring *et al*., 2011), we tested in a standardized infection model (Wölbeling *et al*., 2011) whether human disease isolates are more proficient in causing airway infections than isolates from the environment.

Groups of 20 10 week old female C57BL6J mice were inoculated with 1.5×10^6 colony-forming units (cfu) of the respective *P. aeruginosa* strain into their lower airways by nasal instillation. The course of the infection was then monitored over 144 h by ethological score, body weight, body temperature and lung function. End-points after 6 and 24 h of infection were bacterial cfus, cytokine levels and histology of the murine lungs.

The experiments uncovered a strong interclonal gradient of virulence among the sequenced strains. Mice inoculated with the C40A strain or the B420 outlier

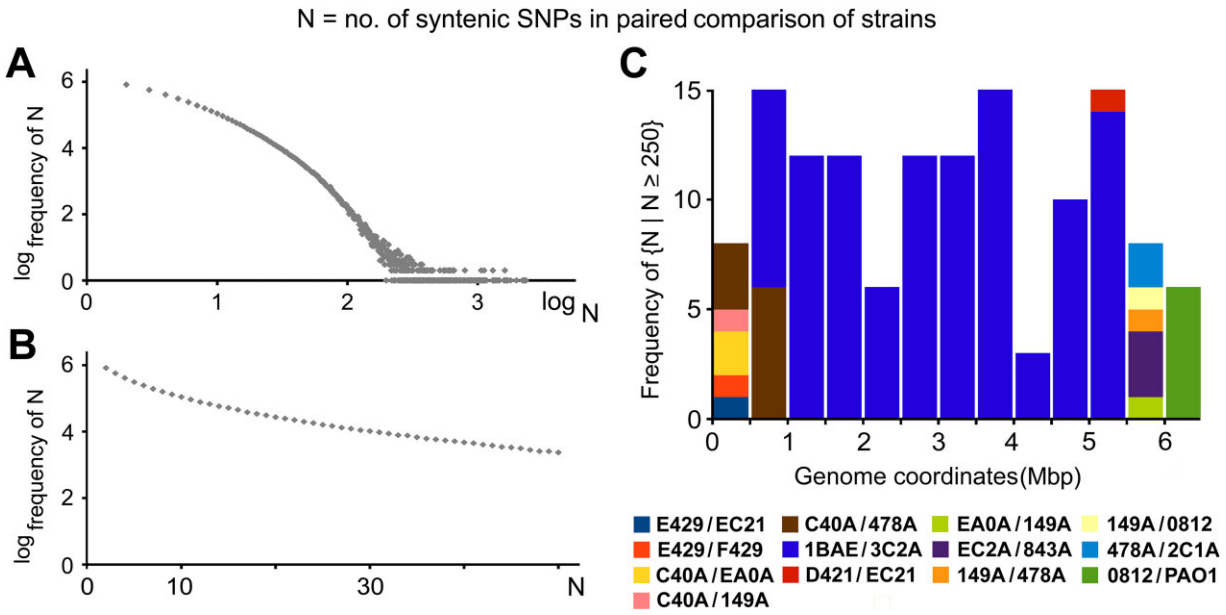


Fig. 6. Syntenic SNPs in the *P. aeruginosa* core genome of the 20 sequenced strains and strain PAO1. The term N is the number of syntenic SNPs in paired comparison of strains. 'N' is equivalent to the term 'haplotype' in case of diploid genomes. Taking the genome coordinates of the SNPs, N was converted into the physical length of a block of identical sequence shared by two genomes.
A, B. Frequency distribution of N_i in dependence of N. The double-logarithmic graph in A shows the complete dataset ranging from a million haplotypes made up of two SNPs to the largest singular haplotypes that consist of more than a thousand syntenic SNPs. The semilogarithmic presentation in B depicts the frequency distribution of common haplotypes made up of 2–50 SNPs.
C. Frequency distribution of the largest blocks of identical sequence ($n \geq 500$) shared by two *P. aeruginosa* genomes along the *P. aeruginosa* chromosome. The genome coordinates of *P. aeruginosa* PAO1 were taken as reference.

behaved like the mock controls. They did not show any clinical signs of infection, had an almost normal lung function during the whole observation period, produced low or no proinflammatory cytokines during infection and exhibited normal lung histology (Fig. 7). The other extreme was observed with the airway isolates of clones OC2E, F429, E429, F469 and the ocean isolate of clone EC21, the latter four grouping with D421 in the smaller genome cluster (see Fig. 5). The same dose as used for C40A or B420 was 100% lethal with these five strains; within 72 h all mice succumbed to death (Fig. 7). The mice did not control the infection as it was indicated by the irreversible decline of body temperature, body weight and lung function, the persistent production of tumour necrosis factor-α (TNFα), interleukin-1ß (IL-1ß) and keratinocyte-derived cytokine (KC) (Fig. S2), and the massive infiltration of neutrophils into the lung (Fig. 7). The bacteria persisted in the lungs and even replicated extracellularly and intracellularly (EC21, OC2E, F469). An intermediate phenotype was seen with the 13 other strains including D421 (Fig. 7). The four more virulent strains in this group killed 11–75% of the infected mice, whereas no lethality was observed with the other nine strains. All survivors showed an intermittent drop of weight, temperature and lung function, which peaked 6–12 h after infection. Body temperature and behaviour normalized by 1–2 days, whereas the recovery of lung function to pre-infection levels required 4–5 days.

In summary, under the conditions of our highly standardized acute murine airway infection model, the full spectrum of possible host responses to *P. aeruginosa* was seen that ranged from unimpaired health to 100% lethality. The pathogenicity of strains did not segregate with habitat. Both environmental and clinical isolates showed the same gradient of virulence in our experimental setting.

Acute lettuce and G. mellonella *infections*

Next, we wanted to know whether similar interclonal gradients of virulence could also be observed with plant and invertebrate hosts. We chose the established *G. mellonella* (Pustelny *et al.*, 2013; Koch *et al.*, 2014; Whiley *et al.*, 2014) and *Lactuca sativa var. longifolia* models (Aendekerk *et al.*, 2005; Gooderham *et al.*, 2009; Starkey and Rahme, 2009; Bielecki *et al.*, 2011) that had been investigated previously in the context of *P. aeruginosa* infections (Fig. 8).

Infections of the last proleg of *G. mellonella* larvae with 5, 10 or 50 cfu of bacteria caused a dose-independent gradient of virulence among the 20 *P. aeruginosa* that differed from the ranking of pathogenicity seen in the mouse experiments (Fig. 9). Most bacterial strains killed

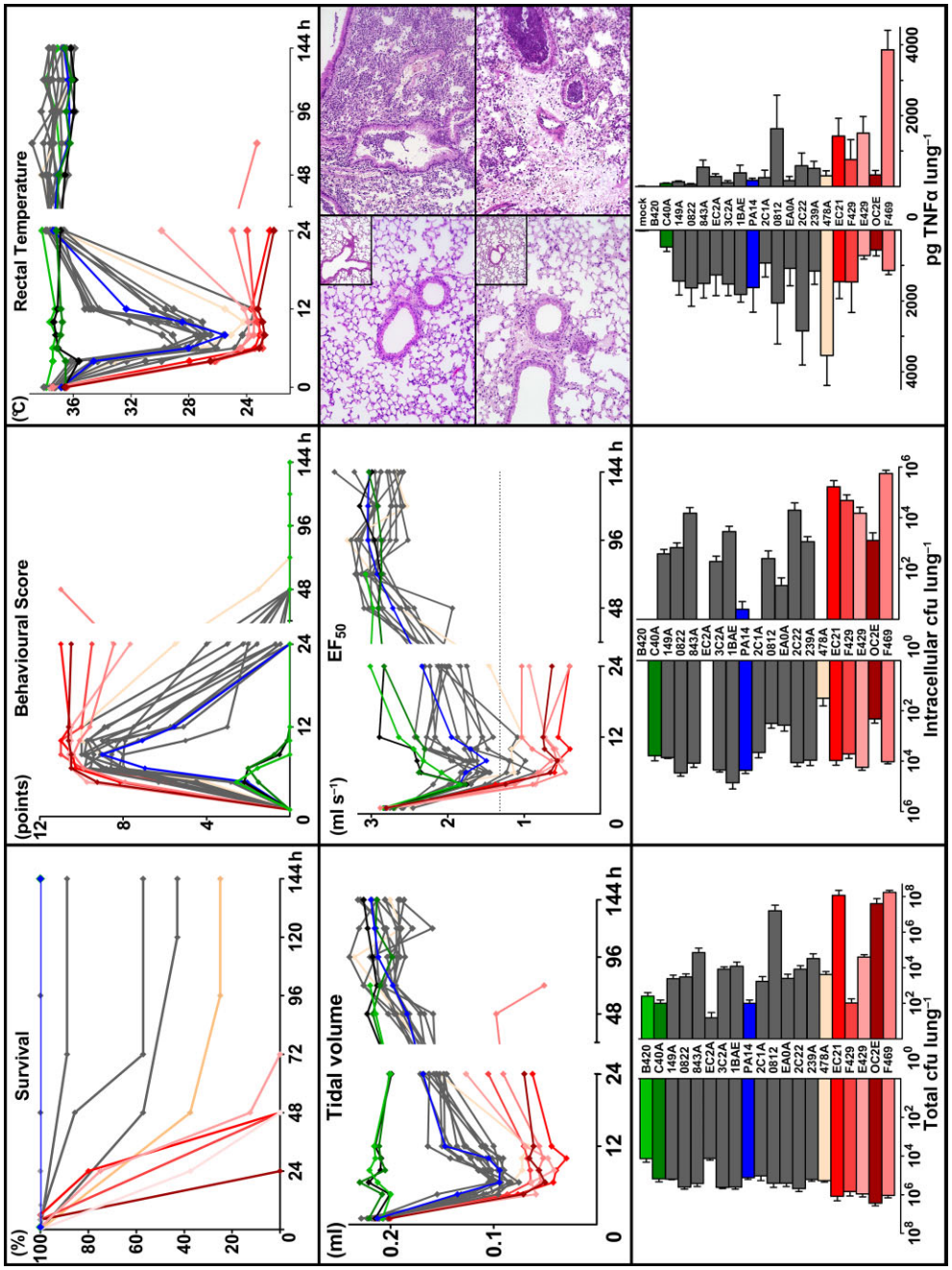


Fig. 7. Airway infection of C57Bl6J mice with the 20 sequenced *P. aeruginosa* strains. Each strain was tested in 20 mice. The six most virulent strains (F469, 0C2E, E429, F429, EC21, 478A) are shown in shades of red, the most avirulent strain B420 in light green and the two most common clones C (genotype: C40A) and PA14 (D421) in dark green and blue respectively. All other strains are marked in grey. The mock control (mice received PBS) is indicated in black. *Upper and middle row:* Kaplan–Meier plots (upper row left) show the survival of mice after nasal instillation of the bacteria (t = 0 h) at 24 h intervals. The infections with the 11 strains with 100% survival of the 20 mice are represented by clone PA14. The time course of behavioural score, rectal temperature and lung function (tidal volume and expiratory flow EF_{50}) is displayed for each bacterial strain by a traverse line of median values of the examined mice. The histological sections (middle row, right) display inflammatory changes in lung tissue 6 h (upper part) and 24 h (lower part) after infection with strain B420 (left) or F469 (right). Inserts present mock infected animals. *Lower row:* Total cfu, intracellular cfu and TNFα levels in murine lungs 6 h (left) and 24 h after infection (right) given as mean and standard deviation (SD) ($n = 6$).

Fig. 8. Gradient of virulence of *P. aeruginosa* in the *G. mellonella* and *L. sativa var. longifolia* infection models. The examples illustrate (from left to right) the phenotype caused by the most virulent, the least pathogenic bacterial strain and the mock control in wax moth larvae (upper row) and lettuce (lower row).

the larvae within the first 24 h, but seven strains moreover exhibited also a slow-killing mode (Fig. 8).

The inoculation of *P. aeruginosa* into the midribs of Romaine lettuce leaves caused rotting of the entire midrib within 2–3 days. Severity of rotting and intensity of brown colouration differed by strain (Fig. 8; Table S4). The group of environmental isolates was more pathogenic than that of the clinical isolates (*U*-test, $P < 0.01$).

The relative pathogenicity of the *P. aeruginosa* strains in the murine, wax moth and lettuce models is visualized

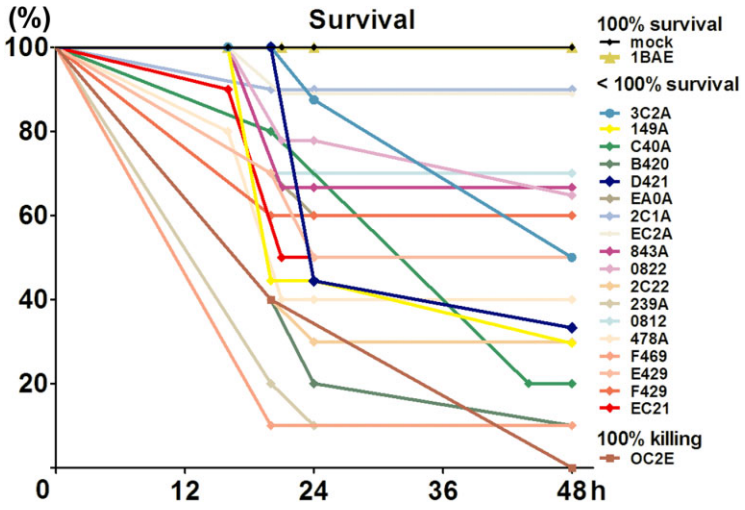


Fig. 9. Kaplan–Meier plot of the survival of *G. mellonella* upon infection with 5 cfu each of the 20 *P. aeruginosa* strains.

Table 2. Rank presentation of the virulence of the 20 sequenced strains in three infection models.

Strain	Rank number in murine airway infection	Rank number in *G. mellonella* larvae infection	Rank number in lettuce leaf infection	Sum of rank number differences between the three models	Sum of rank numbers
C40A	2	18	5	32	25
D421	9	15.5	7	17	31.5
F469	20	15.5	2	36	37.5
0C2E	19	20	17	6	56
E429	18	13	9	18	40
239A	14	19	1	36	34
2C22	13	15.5	6	19	34.5
F429	17	5.5	10	23	32.5
B420	1	15.5	14	29	30.5
EA0A	12	2	3	20	17
0812	11	11.5	8	7	30.5
2C1A	10	1	4	18	15
1BAE	8	4	11	14	23
3C2A	7	10	16	18	33
EC2A	6	3	13	20	22
EC21	16	11.5	19	15	46.5
843A	5	5.5	12	14	22.5
0822	4	9	15	22	28
149A	3	7.5	20	34	30.5
478A	15	7.5	18	21	40.5

Rank numbers were assigned to strains from 1 (lowest virulence) to 20 (highest virulence) according to the observed virulence in the infection models.

in Table 2 by their rank numbers sorted by increasing pathogenicity. The CF lung isolate 0C2E was the only strain of comparably high virulence in all three infection models, and on the other hand the keratitis isolate EA0A and the COPD isolate 2C1A exhibited below-average pathogenicity. The remaining 17 strains had variable rank numbers in the three models. The least pathogenic strains were B420 in mice, 2C1A in the larvae and 239A in the lettuce (Table 2).

The differential responses of the three hosts to the 20 *P. aeruginosa* strains were assessed by the statistical evaluation of rank number differences. The set of bacterial strains evoked a significantly different interclonal gradient of virulence in lettuce than in the two animal models [degrees of freedom (dF) = 2; $\chi^2 = 49.3$; $P < 0.001$]. This statement was even valid at the level of individual strains for the comparison between the murine and the lettuce habitat (dF = 19; $\chi^2 = 34.4$; $P < 0.025$). This statistically verified finding indicated that *P. aeruginosa* orchestrates differential sets of virulence determinants and mechanisms to conquer its many animate habitats.

Clone F469 and B420 genome comparisons

In our lettuce infection experiments, we bypassed the first line of protection, i.e. the physical and chemical barrier of the plant epidermis. In this case, the host–bacterium interaction was determined by the response of the plant to pathogen-associated molecular patterns, siderophores and antimetabolites (Zhang and Zhou, 2010). In contrast to plants, the host–pathogen interaction in mammals is governed by a complex interplay between bacterial fitness and pathogenicity on one hand and innate and adaptive immune responses of the mammalian host on the other. Because among the infections with *P. aeruginosa* the lower airways are the most vulnerable target, we were anxious to analyse the genetic repertoire of the two *P. aeruginosa* strains that exhibited the extremes of lowest and highest pathogenicity in the murine acute airway infection model. The extremes of virulence were represented by the innocuous B420 and the lethal F469 strains (Fig. 7).

B420 is the most common clone among isolates from the inanimate environment (Selezska *et al*., 2012), and F469 has been most frequently detected in isolates from chronic airway infections (Cramer *et al*., 2012). Assuming that the divergent phenotypes are at least partly caused by their genetic blueprints, the B420 and F469 genomes were examined for discriminating features (Table S5). The Venn diagram of the D421, C40A, B420 and F469 genomes (Fig. 3) highlights their combinatorial composition, i.e. all possible fields of singles, dyads, triples and quad are occupied by genes. B420 shares 92.3% of its genes with F469. This large overlap suggests that just a few features of their genetic repertoire may account for their differential pathogenicity.

The annotation of the F469 genome did not readily explain the high virulence of the strain (Table S5; Fig. S3). Notable characteristics of the F469 genome, however,

were the high prevalence of transporters and phage-like DNA and the carriage of unique variants of *popN* (Yang *et al*., 2007) and *pilY1* (Bohn *et al*., 2009). PilY1 and PopN have been shown to modulate bacterial persistence in lung habitats and type III secretion of virulence effectors, respectively, and correspondingly the *popN* and *pilY1* variants in the F469 genome are prime candidates for future experimental work on explaining the heightened pathogenicity of F469.

In contrast to F469, the *in silico* analysis of the B420 genome provided solid evidence why the environmental B420 strain was innocuous in our airway infection model (Tables S4 and S5). B420 lacks all genes for the type III secretion system and its virulence effectors, some genes of the *cupB* and *cupC* fimbrial gene clusters, and the CFTR inhibitory factor *cif*. Hence, the B420 strain is compromised in adhesion to mucosal surfaces and cytotoxicity and unlike most other *P. aeruginosa* cannot perturb the function of mammalian ABC transporters in epithelial cells (Ballok and O'Toole, 2013; Bomberger *et al*., 2014). B420 moreover lacked numerous genes of the RNA, fatty acid and carbohydrate metabolism. In total, 67 genes linked to KEGG pathways were absent in one or more of the 20 sequenced strains, 35 of which were not detected in B420.

Besides the absence of major virulence genes, the comparatively high number of SNPs and indels (Table 1) may contribute to the low pathogenicity of *P. aeruginosa* B420 in airways. The B420 genome harbours the highest number of intragenic insertions and deletions among the 20 sequenced genomes. Interestingly, deleterious frameshift mutations were rare (Table S6). Most out-of-frame mutations were located close to the 3′ end of an ORF and hence may be classified as sequence variants. If the insertion or deletion was located more proximal in the coding sequence, the reading frame was retained by an in-frame deletion or insertion or – more frequently – a frameshift was rescued by a second frameshift in close vicinity (Table S6). Such compensatory frameshifts were typically three to six base pairs apart from each other; the maximal distance was 18 base pairs. These combined frameshifts resulted in amino acid substitutions, deletions of one or two amino acids, or a change of the coding sequence for three to six successive amino acids. The short distance between balanced out-of-frame mutations suggests that in fitness-relevant genes, only those frameshifts persist in the *P. aeruginosa* population that are rescued by no or marginal changes of the amino acid sequence.

Metabolic competence

The redox-active nicotinamide adenine dinucleotides are the 'metabolic currency' of anabolic (NADPH) and catabolic metabolism (NADH) (Fuchs, 1999). It is textbook knowledge that the bacterial cell keeps the ratio of $[NADP^+]/[NADPH]$ low and the ratio of $[NAD^+]/[NADH]$ high. $[NADH]/[(NAD^+) + (NADH)]$ is usually kept near 0.05 (catabolic reduction charge), and $[NADPH]/[(NADP^+) + (NADPH)]$ is kept near 0.5 (anabolic reduction charge) (Fuchs, 1999). We examined the 20 *P. aeruginosa* strains in their levels of the four oxidized or reduced nicotinamide adenine dinucleotides. Unanticipatedly, only a minority of strains exhibited anabolic and catabolic reduction charges in the expected range (Fig. 10). Almost all strains had a higher-than-expected $[NADH]/[(NAD^+) + (NADH)]$ ratio, and half of the strain panel had a lower-than-expected $[NADPH]/[(NADP^+) + (NADPH)]$ ratio. In other words, the homoeostasis of catabolism and anabolism of *P. aeruginosa* under our test conditions was different from the state that is commonly assumed to be valid for bacteria.

The nine most common clones displayed a uniform preponderance of the oxidized state of both NAD and NADP, whereas the less abundant clones and all the environmental isolates but 478A had a preference for the reduced states, particularly NADH (Fig. 10). The pool of phosphorylated forms needed in most biosynthetic pathways only dominated in the most frequent clone C40A. In all other clones, the pools were either balanced (nine strains) or skewed towards the dephosphorylated forms NAD(H). The composition of the nicotinamide adenine dinucleotide pool in the strain panel was variable and did not fit with textbook description for most strains. The most virulent and avirulent strains shared rather similar signatures, indicating that the metabolic competence of a strain was not directly associated with its virulence in our examined infection models. In summary, the known metabolic versatility of *P. aeruginosa* also manifests in a pronounced interclonal diversity of its ratio of anabolic to catabolic reduction charge.

Conclusion

The whole genome comparisons revealed that the major clonal complexes of the *P. aeruginosa* population segregate into outliers and two clusters with the ubiquitous clones C and PA14 as the most prominent representatives. The infections of mice, caterpillars and lettuce with the test strains uncovered an unexpectedly strong interclonal gradient of virulence. Numerous virulence determinants have been discovered in *P. aeruginosa* by the comparison of wild-type and isogenic mutant. The outcome of our highly standardized infection experiments now allows us to sort these pathogenicity factors by relevance. Type III secretion, adhesins and Cif seem to be major stand-alone factors for virulence in mice, whereas the contribution of other elements such as siderophores

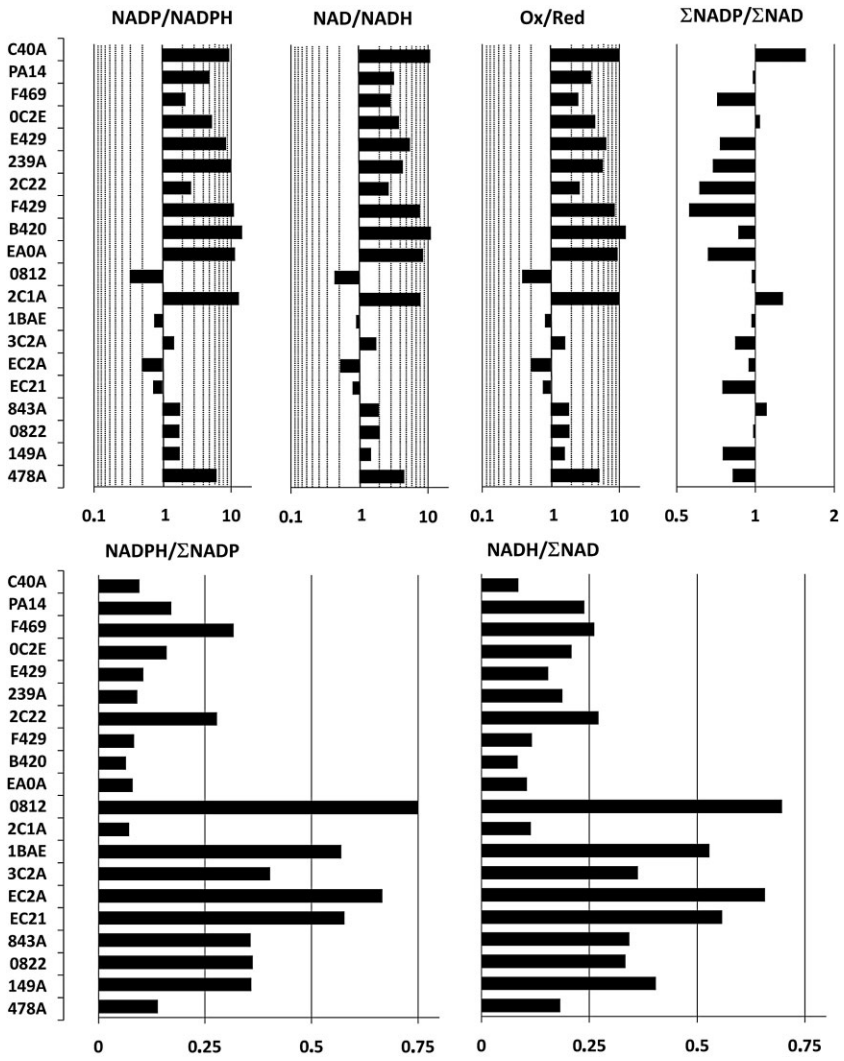


Fig. 10. Ratios of the redox-active nicotinamide adenine dinucleotides in overnight stationary cultures of the 20 *P. aeruginosa* strains. Upper panel (from left to right): $[NADP^+]/[NADPH]$; $[NAD^+]/[NADH]$; $[(NADP^+) + (NAD^+)]/[(NADPH) + (NADH)]$; $[(NADP^+) + (NADPH)]/[(NAD^+) + (NADH)]$. Please note the logarithmic scales at the bottom. Lower panel (from left to right): $[NADPH]/[(NADP^+) + (NADPH)]$; $[NADH]/[(NAD^+) + (NADH)]$.

or type II secretion is combinatorial depending on the asset of pathogenicity factors in the core and accessory genome of the individual strain.

The individual *P. aeruginosa* genome consists of a conserved core, a variable composition of common gene islands and a small set of rare genes that are chiefly hypotheticals of unknown function. Although there are associations between clonal frame and the composition of the accessory genome (Wiehlmann *et al*., 2007), the clonal complexes freely exchange their genes by recombination and transfer of gene islands. The free recombination of the core genome was deduced from the frequency distribution of syntenic SNPs in the sequenced genomes. The key question of molecular epidemiology of whether a bacterial species has a clonal, panmictic or epidemic population structure has been typically investigated in the past by multilocus sequence typing (Maiden, 2006) and/or analysis of polyphasic datasets of phenotypic polymorphisms (Pirnay *et al*., 2009). The outcome was often dependent on the set of parameters selected for analysis. Paired whole genome comparisons of haplotype length allow an unbiased and complete analysis with a definitive outcome, and hence we would like to recommend this approach (see *Experimental procedures*) for any further study on bacterial pangenome and population structure.

Experimental procedures

Pseudomonas aeruginosa *strains*

The sequenced *P. aeruginosa* strains are part of our strain collection that have been genotyped by a custom-made microarray (Wiehlmann *et al*., 2007). The relatedness of strains by multilocus genotype was calculated by the EBURST algorithm (Feil *et al*., 2004). The isolates selected for sequencing belonged to the 15 most common genotypes or to the most frequent genotypes that were yet lacking any clinical isolates (Cramer *et al*., 2012). The strains were deposited at the German Collection of Microorganisms

and Cell Cultures (DSMZ) and are available under accession numbers DSM29238-29241, DSM29272-29281 and DSM29304-29311.

Sequencing

Tagged paired-end and 3 kb mate pair libraries were prepared following the manufacturer's instructions. The 20 genomes were sequenced on an Illumina Genome Analyzer II by GATC-Biotech (Constance, Germany). Sequences were deposited in the European Nucleotide Archive (ENA) hosted by EMBL/EBI, accession no. PRJEB4961.

Alignment of sequencing reads

One hundred one nucleotide paired-end reads for each strain were aligned to the PAO1 reference (NC_002516.2) using the exact and complete alignment software SARUMAN (version 1.0.7; Blom *et al*., 2011) with a maximum of eight mismatches per read and standard Levenshtein distance.

The read datasets of strains F469 and B420 were also mapped to the reference with the alignment software BWA (version BWA-0.6.1; Li and Durbin, 2009) generating 'sam'-formatted alignment files. Unmapped reads were then extracted from the alignment files with an in-house written script and assigned to a not-in-reference read pool of each strain.

Sequence variation analysis

The SARUMAN 'jok' read alignment output was processed with the software READXPLORER (Hilker *et al*., 2014). The read mapping data sets were imported into a READXPLORER project, which includes conversion of the 'jok' output into 'bam' formatted alignment files. Then READXPLORER was used to detect single nucleotide substitutions, deletions and insertions. The examined reference position had to be covered by at least 30 reads of which at least 90% showed the sequence variation. If a read was mapped to more than one site in the reference, quality classification was performed by READXPLORER to assign the read to the best match. Nucleotide variations compared with the reference were extracted from the BWA alignment files by using SAMTOOLS (Li and Durbin, 2009; Li *et al*., 2009). Nucleotide exchanges (SNPs) were filtered from the vcf-formatted SAMTOOLS output files. SNP calls not passing cut-off values for coverage, base call qualities and SNP-call qualities established during former evaluations of Illumina sequencing data (Bezuidt *et al*., 2013) were excluded from the lists. In addition, SNP calls for strains F469 and B420 based on the BWA alignment were cross-checked by comparison with SNP lists generated with READXPLORER from SARUMAN alignments of the same sequencing data.

Predictions of small insertions or deletions (indels) in the F469 or B420 genome compared with the reference sequence were extracted from the SAMTOOLS output, and also from alignments with the STAMPY read mapper (Lunter and Goodson, 2011). Manual inspection of the alignment results for the respective loci, however, unmasked the majority of these predictions as false positives. Indel predictions based on the SARUMAN alignments appeared more reliable but contained false calls or misinterpreted indel events as well. Therefore, for an in depth analysis of indels in F469 and B420, SARUMAN predictions and the top candidates from the BWA/SAMTOOLS and the STAMPY results were assembled in a candidate list. Alignment results for the loci from this list were then inspected with the help of the INTEGRATIVE GENOMICS VIEWER (IGV) (Thorvaldsdóttir *et al*., 2013) for manual verification, modification or dismissal of each indel prediction. Predicted loci were excluded from the list if they were not covered by at least five high-quality sequencing reads and with less than 95% of the reads indicating the indel.

The effects of reliable sequence variations on coding DNA sequences were identified with the program SNPEFF, version 1.9.5 (Cingolani *et al*., 2012).

Absent DNA

PAO1 genomic DNA not present in the F469 or B420 genome was determined by extracting uncovered regions of the reference from the alignment results. Regions eventually prone to low or even no coverage due to an extremely high GC content of more than 80%, which could lower the efficiency of ligation and/or PCR amplification steps during standard Illumina sequencing procedures and for which no 'deletion-spanning' sequence reads could be found, were excluded from the result lists.

De novo *assembly*

De novo assembly of draft genomes was done with the NEWBLER assembler (version 2.8) (Margulies *et al*., 2005). The minimum contig length was set to 150 bases, and the internal read quality trimming of NEWBLER was used. The resulting scaffolds were aligned to the reference strain PAO1 using the contig arrangement software R2CAT (Husemann and Stoye, 2010).

Sequencing reads assigned to the not-in-reference pools after the alignments were assembled to larger contigs with the *de novo* assemblers NEWBLER (version 2.8) (Margulies *et al*., 2005) or VELVET (Zerbino and Birney, 2008). VELVET version 1.2.03 was used with parameters set to a minimum read coverage of five and a kmer-size of 31. For F469 and B420 analysis, both assemblers were used in parallel, and contigs from both programs were merged in case of complementary results.

Accessory genome analysis

The assembled contigs, representing the accessory genome of the analysed strains, were analysed by blastx comparisons against the UniProt database (Apweiler *et al*., 2004) in order to detect known genes from other *P. aeruginosa* genomes or genes from other bacterial species in the accessory genomes. The contigs were also used for the detection of known *P. aeruginosa* GIs or RGPs by doing blastn comparisons against these sequences. Blastn results were used to determine presence and degree of conservation of these elements in the F469 and B420 accessory genomes. In a second approach, paired-end Illumina reads were aligned

to known GI and RGPs using BWA (Li and Durbin, 2009) with default parameters. Uncovered reference regions were identified using the genomeCoverageBed utility from BEDTOOLS (Quinlan and Hall, 2010). Coverage percentages were calculated for each strain using an in-house Perl script. The coverage percentage tables then were used to create the coverage heatmaps using the gplots R package.

Pan and core genome analysis

Draft genomes were created from the NEWBLER *de novo* assemblies as described above. All not-in-reference scaffolds were appended at the end of the ordered scaffold list. To prevent gene prediction across scaffold borders, the whole list of scaffolds was concatenated by a stop linker on all six reading frames (CTAGCTAGCTAG) using an in-house written script. Automatic annotation of the draft genomes was done in GENDB (version 2) (Meyer *et al*., 2003) using PRODIGAL (version 2.6) (Hyatt *et al*., 2010) for gene predictions. The annotated draft genomes were exported as GenBank formatted files from GENDB. Afterwards the GenBank files of the draft genomes and the reference strain PAO1 were used to create an EDGAR (version 1.2) (Blom *et al*., 2009) project for the pangenome and core genome analysis. The score ratio value (SRV) (Lerat *et al*., 2003) used as master cut-off for this project is 30%. This means that only reciprocal BLASTP hits for coding sequences with an SRV higher than or equal to 30% are marked as being present in two compared genomes. Based on this parameter, the pangenome and core genome were calculated with EDGAR. Additionally, a phylogenetic tree based on the core genome, a list of singleton genes, which only occur in one of the genomes, and a list of all pangenome genes were created with EDGAR.

Syntenic SNPs

Because according to our knowledge the procedure to determine the number of syntenic SNPs in paired genome comparison of bacterial strains has yet not been described, the analysis is explained in more detail. To simplify the understanding for the reader, the entity N, i.e. the number of syntenic SNPs in paired comparison of two bacterial genomes, is called a haplotype in analogy to the usage of this term for diploid genomes.

First, the RGP-free PAO1 core genome sequence was taken as a reference to identify SNPs in the 20 sequenced genomes. To construct haplotypes, a matrix was constructed that contained columns of all 192 443 quality-controlled SNPs ordered by genome position and rows of all strains (PAO1 reference and the 20 sequenced strains). The value 0 was assigned to nucleotides that match with the PAO1 reference, and the value 1 was assigned to the nucleotide substitution. Next, all 210 possible combinations of two genomes were compared in their similarity of SNP pattern. Haplotypes were identified by counting successive matches of binary pattern until the first mismatch. The number N of syntenic SNPs was then inserted at each SNP position of the haplotype. For this purpose, a second matrix of haplotypes was constructed that consisted of all SNP positions as columns, the 210 paired comparisons as rows and the numbers of syntenic SNPs, i.e. the haplotypes, as entries. This matrix was used to extract the haplotypes of the 210 paired comparisons, to determine large haplotypes above a floating threshold and to evaluate the frequency distribution of haplotype size in the strain panel. The distribution of haplotype size along the chromosome was calculated in sliding windows of 21 successive SNPs. Automatic analysis was performed with in-house Perl scripts.

Assays of nicotinamide adenine dinucleotides

Strains were precultured in 5 ml of tryptic soy broth (TSB) for 4 h at 37°C. Two hundred microlitres thereof were inoculated into 20 ml TSB and incubated for 12 h at 37°C and 200 r.p.m. Bacteria were pelleted and lysed with acid or base. Purified supernatants were subjected to enzyme cycling-based colourimetric assay to determine the amounts of NAD^+, NADH, $NADP^+$ and NADPH according to the protocol described by Kern and colleagues (2014).

Infection experiments

Mice. Ten to twelve week old female C57BL6J mice (Charles River Germany) were maintained in microisolator cages with filter top lids at 21 ± 2°C, 50% ± 5% humidity and a 14/10 h light–dark cycle. They were supplied with autoclaved, acidulated water and fed *ad libitum* with autoclaved standard diet. All animal procedures were approved by the local animal welfare committee and carried out according to the guidelines of the German regulations for animal protection.

Culturing of bacteria. Strains of the culture collection were streaked on tryptic soy agar plates and incubated at 37°C for 14 h. Colonies were then inoculated into TSB to a final optical density of 0.225 at 550 nm. The bacteria were cultured at 37°C for 1 h with shaking, harvested by centrifugation, washed with HEPES/saline and then re-suspended in Hepes/saline at a density of 1.0×10^8 cfu ml^{-1}. To adjust for the up to eightfold different growth rates of the strains, the factors of dilution were calculated from growth curves of the strains that had been recorded in prior experiments. This standardized procedure was applied to all infection experiments.

Murine infection protocol (Fig. 11). Twenty 10–12 week old female C57Bl/6J mice were anaesthetized (midazolam/ketamin i.p.) and inoculated with 1.5×10^6 cfu of each *P. aeruginosa* strain via intranasal instillation. End-point sampling was performed at 6 and 24 h p.i. on six mice each. Lungs were weighed and divided into pieces which were weighed again. The cranial lobe and the middle lobe of the right lung were put into paraformaldehyde for histopathology; the caudal lobe and the accessory lobe were shock frozen for cytokine analysis and the left lung was used for the determination of total and intracellular cfu. The course of the infection in the other eight mice was followed for 144 h by regular assessment (4, 6, 8, 10, 12, 24, 48, 72, 96, 120 and 144 h) of the behavioural score (Munder *et al*., 2005), rectal temperature, body weight and non-invasive headout spirometry.

Murine lung function. Non-invasive head-out spirometry investigating 14 lung function parameters was performed on

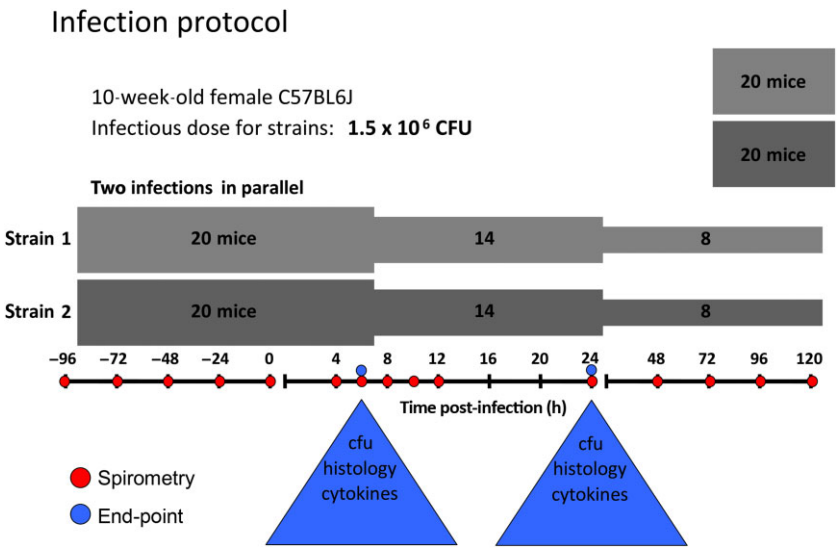


Fig. 11. Protocol of the acute murine airway infection experiments with the *P. aeruginosa* strains.

conscious restrained mice (Wölbeling *et al.*, 2010). In brief, mice were restrained in glass inserts with a set of membranes around their neck. Respiration causes air to flow through a pneumotachograph positioned above the thorax of the mice. A pressure transducer creates an electrical signal, which is analysed using special software (NOTOCORD HEM, Version 4.2.0.241, Notocord Systems SAS, Croissy Sur Seine, France). Spirometry hardware was supplied by Hugo Sachs Elektronik – Harvard Apparatus (March-Hugstetten, Germany). The parameters of tidal volume (measured in ml) and the flow at 50% of the expiratory tidal volume (EF50, measured in millilitres per second) were selected to characterize murine lung function during infection. Data were analysed by GRAPHPAD PRISM software (Version 5.0, GraphPad Software, San Diego, CA, USA). Median values and standard deviation were calculated. Mice became accustomed to spirometry by daily training during the 96 h period prior to infection.

Total and intracellular bacterial cfus

The left lungs of the euthanized mice were ligated, resected and homogenized (Polytron PT 1200 Homogenizer, Germany). Total bacterial numbers were assessed from homogenized lung tissue, which was additionally lysed for 10 min in a saponin solution (5 mg ml^{-1}) to release intracellular bacteria. Serial dilutions of the homogenates were cultured on Luria–Bertani plates using the drop plate method (Herigstad *et al.*, 2001). For determination of the intracellular bacteria, lung pieces were incubated with polymyxin (100 μg ml^{-1}) for 1 h, washed three times with PBS and then processed as described for the determination of total cfus. cfus were normalized for lung weight.

Cytokine analysis

The shock-frozen lung pieces were homogenized in 400 μl PBS, and aliquots were used directly in commercial ELISA assays (Roche or R&D). The ELISA measurements of cytokines TNFα, IL-1β and KC were performed according to the protocol of the vendor. Cytokine concentrations were calculated by comparison with standard curves and were then normalized for lung weight or lung protein the latter determined by the Bradford method.

Galleria mellonella *infection protocol*

Bacteria grown in TSB were adjusted to 2.5×10^2, 5.0×10^2 and 2.5×10^3 cfu ml^{-1} physiological saline. Ten *G. mellonella* larvae per strain and dose were inoculated with 20 μl containing 5, 10 and 50 cfu respectively. Bacteria were injected into the last proleg of the larvae (Pustelny *et al.*, 2013). Survival of the larvae at 37°C was monitored for 72 h. Dead larvae did not respond anymore to tapping and/or turned black because of melanization. Untreated larvae were used for controlling the culturing conditions. Larvae injected with isotonic NaCl solution served as mock control.

Lettuce infection protocol

Leaves of freshly harvested *L. sativa var. longifolia* were cut from the core, washed with 0.1% (v/v) bleach and rinsed twice with distilled water (Starkey and Rahme, 2009). Leaves were placed into 14 cm diameter Petri dishes onto Whatman paper soaked with 10 mM $MgCl_2$ solution. Ten microlitres of 10 mM $MgCl_2$ solution containing 10^4 or 10^6 cfu *P. aeruginosa* were instilled into the midrib of the lettuce leaf at least 3 cm apart from cutting edge. Petri dishes were sealed and incubated at 37°C for up to 72 h. The progress of infection was documented by protocol, i.e. visual recording and photographs taken always under identical conditions. Completely infected leaves were marked. The size of the infected area at the 44 h and 52 h time points was quantified by computer-assisted planimetry.

Statistics

The outcome of the infection experiments in lettuce, insect larvae and mice was compared by rank tests (Weber, 1967).

Acknowledgements

The authors would like to thank S. Wiehlmann, A. Burhop and A. Zheng for excellent technical assistance and A. Bragonzi (Milan), T. Köhler (Geneva), K. Kogure (Tokyo), A. Kumar (New Delhi), T.F. Murphy (Buffalo), J. Sikorski (Braunschweig) and C. Winstanley (Liverpool) for the provision of strains. This work was supported by grants of the Bundesministerium für Bildung und Forschung (programme 'Medical Infection Genomics', 0315827A) and of the Deutsche Forschungsgemeinschaft (SFB 900, project A2).

References

Aendekerk, S., Diggle, S.P., Song, Z., Høiby, N., Cornelis, P., Williams, P., and Cámara, M. (2005) The MexGHI-OpmD multidrug efflux pump controls growth, antibiotic susceptibility and virulence in *Pseudomonas aeruginosa* via 4-quinolone-dependent cell-to-cell communication. *Microbiology* **151:** 1113–1125.

Apweiler, R., Bairoch, A., and Wu, C.H. (2004) Protein sequence databases. *Curr Opin Chem Biol* **8:** 76–80.

Ballok, A.E., and O'Toole, G.A. (2013) Pouring salt on a wound: *Pseudomonas aeruginosa* virulence factors alter Na+ and Cl- flux in the lung. *J Bacteriol* **195:** 4013–4019.

Bezuidt, O.K., Klockgether, J., Elsen, S., Attree, I., Davenport, C.F., and Tümmler, B. (2013) Intraclonal genome diversity of *Pseudomonas aeruginosa* clones CHA and TB. *BMC Genomics* **14:** 416.

Bielecki, P., Puchałka, J., Wos-Oxley, M.L., Loessner, H., Glik, J., Kawecki, M., *et al*. (2011) In-vivo expression profiling of *Pseudomonas aeruginosa* infections reveals niche-specific and strain-independent transcriptional programs. *PLoS ONE* **6:** e24235.

Blom, J., Albaum, S.P., Doppmeier, D., Pühler, A., Vorhölter, F.J., Zakrzewski, M., and Goesmann, A. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* **10:** 154.

Blom, J., Jakobi, T., Doppmeier, D., Jaenicke, S., Kalinowski, J., Stoye, J., and Goesmann, A. (2011) Exact and complete short-read alignment to microbial genomes using graphics processing unit programming. *Bioinformatics* **27:** 1351–1358.

Bohn, Y.S., Brandes, G., Rakhimova, E., Horatzek, S., Salunkhe, P., Munder, A., *et al*. (2009) Multiple roles of *Pseudomonas aeruginosa* TBCF10839 PilY1 in motility, transport and infection. *Mol Microbiol* **71:** 730–747.

Bomberger, J.M., Ely, K.H., Bangia, N., Ye, S., Green, K.A., Green, W.R., *et al*. (2014) *Pseudomonas aeruginosa* Cif enhances the ubiquitination and proteasomal degradation of the transporter associated with antigen processing (TAP) and reduces MHC class I antigen presentation. *J Biol Chem* **289:** 152–162.

Bondy-Denomy, J., Pawluk, A., Maxwell, K.L., and Davidson, A.R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493:** 429–432.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6:** 80–92.

Cramer, N., Wiehlmann, L., Ciofu, O., Tamm, S., Høiby, N., and Tümmler, B. (2012) Molecular epidemiology of chronic *Pseudomonas aeruginosa* airway infections in cystic fibrosis. *PLoS ONE* **7:** e50731.

Döring, G., Parameswaran, I.G., and Murphy, T.F. (2011) Differential adaptation of microbial pathogens to airways of patients with cystic fibrosis and chronic obstructive pulmonary disease. *FEMS Microbiol Rev* **35:** 124–146.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., and Spratt, B.G. (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186:** 1518–1530.

Fuchs, G. (1999) Basic prerequisites for cellular life. In *Biology of the Prokaryotes*. Lengeler, J.W., Drews, G., and Schlegel, H.G. (eds). Stuttgart, Germany: Thieme, pp. 110–160.

Gansner, E.R., and North, S.C. (2000) An open graph visualization system and its applications to software engineering. *Software Pract Exper* **30:** 1203–1233.

Gansner, E.R., Koren, Y., and North, S.C. (2005) Graph drawing by stress majorization. *Lect Notes Comput Sc* **3383:** 239–250.

Gooderham, W.J., Gellatly, S.L., Sanschagrin, F., McPhee, J.B., Bains, M., Cosseau, C., *et al*. (2009) The sensor kinase PhoQ mediates virulence in *Pseudomonas aeruginosa*. *Microbiology* **155:** 699–711.

Herigstad, B., Hamilton, M., and Heersink, J. (2001) How to optimize the drop plate method for enumerating bacteria. *J Microbiol Methods* **44:** 121–129.

Hilker, R., Stadermann, K.B., Doppmeier, D., Kalinowski, J., Stoye, J., Straube, J., *et al*. (2014) ReadXplorer-visualization and analysis of mapped sequences. *Bioinformatics* **30:** 2247–2254.

Husemann, P., and Stoye, J. (2010) r2cat: synteny plots and comparative assembly. *Bioinformatics* **26:** 570–571.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11:** 119.

Kern, S.E., Price-Whelan, A., and Newman, D.K. (2014) Extraction and measurement of $NAD(P)^+$ and NAD(P)H. *Methods Mol Biol* **1149:** 311–323.

Klockgether, J., Würdemann, D., Reva, O., Wiehlmann, L., and Tümmler, B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J Bacteriol* **189:** 2443–2459.

Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C.F., and Tümmler, B. (2011) *Pseudomonas aeruginosa* genomic structure and diversity. *Front Microbiol* **2:** 150.

Koch, G., Nadal-Jimenez, P., Cool, R.H., and Quax, W.J. (2014) Assessing *Pseudomonas* virulence with nonmammalian host: *Galleria mellonella*. *Methods Mol Biol* **1149:** 681–688.

Koonin, E.V., and Wolf, Y.I. (2012) Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front Cell Infect Microbiol* **2:** 119.

Lee, D.G., Urbach, J.M., Wu, G., Liberati, N.T., Feinbaum, R.L., Miyata, S., *et al*. (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol* **7:** R90.

Lee, X., Fox, A., Sufrin, J., Henry, H., Majcherczyk, P., Haas, D., and Reimmann, C. (2010) Identification of the biosynthetic gene cluster for the *Pseudomonas aeruginosa* antimetabolite L-2-amino-4-methoxy-trans-3-butenoic acid. *J Bacteriol* **192:** 4251–4255.

Lerat, E., Daubin, V., and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biol* **1:** E19.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Lunter, G., and Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* **21:** 936–939.

Maiden, M.C. (2006) Multilocus sequence typing of bacteria. *Annu Rev Microbiol* **60:** 561–588.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J.M., Koehrsen, M., *et al*. (2008) Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci USA* **105:** 3100–3105.

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., *et al*. (2003) GenDB – an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* **31:** 2187–2195.

Munder, A., Zelmer, A., Schmiedl, A., Dittmar, K.E., Rohde, M., Dorsch, M., *et al*. (2005) Murine pulmonary infection with *Listeria monocytogenes*: differential susceptibility of BALB/c, C57BL/6 and DBA/2 mice. *Microbes Infect* **7:** 600–611.

Pirnay, J.P., Bilocq, F., Pot, B., Cornelis, P., Zizi, M., Van Eldere, J., *et al*. (2009) *Pseudomonas aeruginosa* population structure revisited. *PLoS ONE* **4:** e7740.

Pustelny, C., Brouwer, S., Müsken, M., Bielecka, A., Dötsch, A., Nimtz, M., *et al*. (2013) The peptide chain release factor methyltransferase PrmC is essential for pathogenicity and environmental adaptation of *Pseudomonas aeruginosa* PA14. *Environ Microbiol* **15:** 597–609.

Quinlan, A.R., and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

Rahme, L.G., Stevens, E.J., Wolfort, S.F., Shao, J., Tompkins, R.G., and Ausubel, F.M. (1995) Common virulence factors for bacterial pathogenicity in plants and animals. *Science* **268:** 1899–1902.

Ramos, J.L. (ed.) (2004–2010a) *Pseudomonas*. Vol. 1–3. New York, USA: Plenum Press.

Ramos, J.L. (ed.) (2004–2010b) *Pseudomonas*. Vol. 4–6. Heidelberg, Germany: Springer.

Römling, U., Kader, A., Sriramulu, D.D., Simm, R., and Kronvall, G. (2005) Worldwide distribution of *Pseudomonas aeruginosa* clone C strains in the aquatic environment and cystic fibrosis patients. *Environ Microbiol* **7:** 1029–1038.

Selezska, K., Kazmierczak, M., Müsken, M., Garbe, J., Schobert, M., Häussler, S., *et al*. (2012) *Pseudomonas aeruginosa* population structure revisited under environmental focus: impact of water quality and phage pressure. *Environ Microbiol* **14:** 1952–1967.

Starkey, M., and Rahme, L.G. (2009) Modeling *Pseudomonas aeruginosa* pathogenesis in plant hosts. *Nat Protoc* **4:** 117–124.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., *et al*. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406:** 959–964.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192.

Weber, E. (1967) *Grundriß der Biologischen Statistik*, 6th edn. Jena, Germany: VEB Gustav Fischer Verlag.

Whiley, R.A., Sheikh, N.P., Mushtaq, N., Hagi-Pavli, E., Personne, Y., Javaid, D., and Waite, R.D. (2014) Differential potentiation of the virulence of the *Pseudomonas aeruginosa* cystic fibrosis liverpool epidemic strain by oral commensal Streptococci. *J Infect Dis* **209:** 769–780.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G., *et al*. (2007) Population structure of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA* **104:** 8101–8106.

Winsor, G.L., Lam, D.K., Fleming, L., Lo, R., Whiteside, M.D., Yu, N.Y., *et al*. (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res* **39** (Database issue): D596–D600.

Winstanley, C., Langille, M.G., Fothergill, J.L., Kukavica-Ibrulj, I., Paradis-Bleau, C., Sanschagrin, F., *et al*. (2009) Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res* **19:** 12–23.

Wölbeling, F., Munder, A., Stanke, F., Tümmler, B., and Baumann, U. (2010) Head-out spirometry accurately monitors the course of *Pseudomonas aeruginosa* lung infection in mice. *Respiration* **80:** 340–346.

Wölbeling, F., Munder, A., Kerber-Momot, T., Neumann, D., Hennig, C., Hansen, G., *et al*. (2011) Lung function and inflammation during murine *Pseudomonas aeruginosa* airway infection. *Immunobiology* **216:** 901–908.

Yang, H., Shan, Z., Kim, J., Wu, W., Lian, W., Zeng, L., *et al*. (2007) Regulatory role of PopN and its interacting partners in type III secretion of *Pseudomonas aeruginosa*. *J Bacteriol* **189:** 2599–2609.

Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Zhang, J., and Zhou, J.M. (2010) Plant immunity triggered by microbial molecular signatures. *Mol Plant* **3:** 783–793.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. The figure shows the 210 boxplots of the frequency distribution of syntenic SNPs, i.e. boxplots of haplotype length, derived from the 210 paired comparisons of syntenic SNPs in the panel of 20 sequenced strains and strain PAO1.
Fig. S2. The blocks indicate the levels [mean ± standard deviation (SD), $n = 6$] of interleukin-1ß and keratinocyte-derived cytokine in murine lungs 6 and 24 h after infection with the different *P. aeruginosa* strains.
Fig. S3. The pie charts summarize the composition of the accessory genome of F469 and B420 in terms of functional class of protein (A) and the taxonomic position of the most closely related orthologue (B).
Table S1. The table with the heading 'The *P. aeruginosa* pangenome' lists the 13 527 distinct genes that were annotated in the panel of PAO1 and the 20 sequenced strains. The table with the heading 'ORFs specific for single *P. aeruginosa* strains' lists all predicted strain-specific genes.
Table S2A. The tables list all detected sequence variations for each of the 20 sequenced isolates. Because of size limitations, the tables had to be allocated to two separate files. Table_S2A: clones C40A, D421, F469, 0C2E, E429, 239A, 2C22, B420.
Table S2B. Table _S2B: clones F429, EA0A, 0812, 2C1A, 1BAE, 3C2A, EC2A, EC21, 843A, 0822, 149A, 478A, SNP statistics.
Table S3. The table titled 'Large blocks of identical sequence shared by two isolates' lists the length and genome coordinates of all syntenic SNP contigs ('haplotype blocks') longer than 20 kb.
Table S4. The table lists the infected area of lettuce leaves 52 h after infection with *P. aeruginosa*.
Table S5. The table lists all SNPs detected in the B420 and F469 genomes that affect start or stop codons or cause non-conservative amino acid substitutions compared to the PAO1 reference.
Table S6. Genome features of F469 and B420 (genomic islands, absent PAO1 loci, intragenic indels).

Chapter 5

Intraclonal genome diversity of the major *P. aeruginosa* clones C and PA14

5.1 Background

As we described in Chapter 4 *P. aeruginosa* is a ubiquitous environmental organism which can be found in diverse ecological niches. Moreover, this bacterium is an opportunistic pathogen capable of infecting plants, animals and humans. As a human opportunistic pathogen, *P. aeruginosa* has become one of the leading causes of mortality among Cystic Fibrosis patients as well as one of the main causes of hospital-associated pneumonia. Its ecological adaptability is based on a broad genetic repertoire and phenotypic adaptation. In Chapter 4 we already demonstrated that the *P. aeruginosa* pangenome is made up of a core genome of about 4,000 genes common to all *P. aeruginosa*, a flexible accessory genome of 10,000 genes and ten thousands of genes only present in a few strains[81]. This genetic variability provides different grades of virulence, from moderately to highly virulent strains.

Clones C and PA14 are the worldwide most abundant clonal complexes in the *P. aeruginosa* population.

Reference strain PA14 is a highly virulent, its genome published in 2004 contains two pathogenicity islands which carry several genes implicated in virulence[82-84].

Clone C strains infect patients worldwide, it has been found in the clinical as well as aquatic environments[85-87].

5.2 About the manuscript

In the manuscript intraclonal genome diversity was studied in hundred isolates of the predominant *P. aeruginosa* clones C and PA14. They have been selected from the environment, and from both acute and chronic infections.

The analysis of the core and accessory genome revealed a highly conserved core genome and a highly variable accessory genome within the clone.

In order to evaluate the impact of recombination processes in Clone C and Clone PA14 the novel method described in Chapter one was applied. The length of syntenic segments of both clones was calculated. The results show that large blocks of identical sequences are shared within the clone. The median size of shared blocks was 99 kb in clone C and 163 kb in clone PA14. Haplotypes were used to visualize the relatedness of strains in a split-tree.

Author's contribution.

Sebastian Fischer and Nina Cramer conducted the study.

I performed the intraclonal recombination analysis of P. aeruginosa clones C and PA14 as well as the graphs and tables describing the recombination process.

For Tables and Supplementary material, please refer to the DVD attached to the thesis.

Intraclonal genome diversity of the major *Pseudomonas aeruginosa* clones C and PA14

Sebastian Fischer,[1†] Nina Cramer,[1†] Patricia Moran Losada,[1] Philippe Chouvarine,[1] Sarah Dethlefsen,[1] Colin F. Davenport,[1] Marie Dorda,[1] Alexander Goesmann,[2] Rolf Hilker,[2] Samira Mielke,[1] Torben Schönfelder,[1] Sebastian Suerbaum,[3] Oliver Türk,[1] Sabrina Woltemate,[3] Lutz Wiehlmann,[1] Jens Klockgether[1] and Burkhard Tümmler[1,4*]

[1]Clinical Research Group 'Molecular Pathology of Cystic Fibrosis and Pseudomonas Genomics', OE 6710, Hannover Medical School, Hannover, Germany.

[2]Bioinformatics and Systems Biology, Justus-Liebig-Universität, Gießen, Germany.

[3]Institute for Medical Microbiology and Hospital Epidemiology, OE 5210, Hannover Medical School, Hannover, Germany.

[4]Biomedical Research in Endstage and Obstructive Lung Disease (BREATH), German Center for Lung Research, Hannover, Germany.

Running title: Genome Diversity of two major *P. aeruginosa* clones

*For correspondence: Burkhard Tümmler, Clinical Research Group, Clinic for Paediatric Pneumology, Allergology and Neonatology, OE 6710, Hannover Medical School, Carl-Neuberg-Str. 1, D-30625 Hannover, Germany, phone: +49-511-5322920, Fax: +49-511-5326723; email: tuemmler.burkhard@mh-hannover.de

[†]These authors contributed equally to this work.

Summary

Bacterial populations differentiate at the subspecies level into clonal complexes. Intraclonal genome diversity was studied in hundred isolates of two dominant *Pseudomonas aeruginosa* clones collected from the inanimate environment, acute and chronic infections. The core genome was highly conserved among clone members with a sequence diversity of 10^{-6}, but the composition of the accessory genome was as variable within the clone as between unrelated clones and each strain carried a large cargo of unique genes. Within-clone genome diversity translated into up to 10,000-fold different rates of growth and persistence in clonal communities demonstrating that few genomic differences among natural isolates are sufficient to generate large gradients of competitive fitness.

Introduction

Within a bacterial species the individual strains typically segregate into distinct clonal complexes that share more genetic and phenotypic features among themselves than with clonally unrelated strains (Robinson et al., 2010). Correspondingly, genome diversity is significantly lower within clones than between clones. Here we report on the intraclonal genome diversity of the two most common clones in the global population of the opportunistic pathogen *Pseudomonas aeruginosa* (Wiehlmann et al., 2007).

P. aeruginosa is a ubiquitous and metabolically versatile Gram-negative bacterium that thrives in soil and aquatic habitats and colonizes the animate surfaces of plants, animals and humans (Ramos et al., 2004-2015). *P. aeruginosa* has become one of the most common causative agents of acute or chronic infections in predisposed or immunocompromised hosts worldwide (American Thoracic Society, 2005). Nosocomial infections are associated with substantial morbidity and mortality, for example, *P. aeruginosa* pneumonia and sepsis have a 20% – 60% lethality (American Thoracic Society, 2005).

The success of *P. aeruginosa* as a cosmopolitan aquatic bacterium and opportunistic pathogen is based on its broad genetic repertoire. Its pangenome consists of a core of about 4,000 genes common to all *P. aeruginosa*, a flexible accessory genome of 10,000 genes and at least a further hundred thousand genes only present in a few clones or strains (Hilker et al., 2015). We have selected 100 isolates of the predominant *P. aeruginosa* clones C and PA14 from the inanimate environment, acute and chronic infections to study general features of intraclonal genome variation, i.e. the conservation and motility of core and accessory genome, the nature and frequency of sequence variation and mutation and the gradient of fitness to grow and to persist within a community of clonal strains.

Results

The NN2 clone C genome

Clone C is the most abundant clone in the worldwide *P. aeruginosa* population (Römling et al., 1994; Wiehlmann et al., 2007). Completely sequenced *P. aeruginosa* genomes are already available for several clones but not clone C. Hence we decided to completely sequence a phenotypically characterized clone C strain by DNA- and RNA-seq. The cystic fibrosis (CF) isolate NN2 was selected as the reference strain for clone C because strain NN2 was the first *P. aeruginosa* clone C isolate in a *P. aeruginosa* naive CF subject and its subsequent genomic microevolution in the CF host for the next 25 years is known (Cramer et al., 2011). Moreover, phenotypic traits of morphotype, motility, virulence, fitness (Cramer et al., 2011), a physical genome map (Schmidt et al., 1996) and the sequence of genomic islands (Larbig et al., 2002; Klockgether et al., 2004) have been investigated before.

The 6,902,967 bp large NN2 genome encodes 6,601 open reading frames and at least 557 non-coding RNAs (Fig 1a, Table 1, Table S1). Strain NN2 shares 5,379 orthologs with reference strain PAO1 and harbours a repertoire of 48 islands in its accessory genome (Klockgether et al., 2011). The majority of the 1,223 non-PAO1 ORFs encode proteins of unknown function indicating that NN2 carries a large yet unexplored set of clone- or strain-specific genes. Reliably annotated non-PAO1 ORFs code for elements of horizontal gene transfer, maintenance and defense of genomic integrity, enzymes, transcriptional regulators and heavy metal resistance proteins. Instructive examples are multiple copies of DNA repair genes or condensins not yet reported in *P. aeruginosa* that assist in the correct folding of the chromosome (Badrinarayanan et al., 2012; She et al., 2013). These extra features may confer clone-specific fitness traits to clone C so that it could become the most abundant clone in both environmental and disease habitats (Wiehlmann et al., 2007).

The presence of non-coding (nc) RNAs was explored by RNA-seq of the transcriptome profiles of planktonic NN2 bacteria grown in fermenters in nutrient-rich tryptic soy broth until mid-exponential or early stationary phase, respectively. Besides known intergenic ncRNAs (Gómez-Lozano et al., 2012) we detected further 39 intergenic and – for the first time – 288 intragenic ncRNAs in the NN2 transcriptomes (Table S1). The intragenic ncRNAs consist of a small population of short 51 – 87 nt transcripts (n = 46, median 66 nt) and a larger population with a Gaussian distribution of transcript length (n = 242, range 92 – 1034 nt, median 257 nt). Two to four ncRNAs were found within 42 operons and 28 gene loci. Intragenic ncRNAs were overrepresented in genes encoding elements of motility, metabolic enzymes and DNA-associated processes and underrepresented among transcriptional regulators.

The transcriptome data also uncovered the position of the transcriptional start sites of NN2 mRNA transcripts (Table S2). Compared to the PA14 transcriptome (Wurtzel et al., 2012), strain NN2 synthesized fewer leaderless transcripts, but a larger portion of short 5'-UTRs of 10 – 17 bp in length. The length distribution was indistinguishable between the two strains for 5'-UTRs longer than 33 bp that made up 60 % of expressed genes (Fig 1b). The 250 bp region upstream of these genes is enriched for tetranucleotides with low base-pair stacking energy (Fig 1c) consistent with its role to initiate transcription from single-stranded DNA.

Genome diversity of clone C and clone PA14 strains

Having complete genome sequences of the reference strains PA14 (Lee et al., 2006) and NN2 at hand (this work), we next explored the intraclonal diversity of these two major clones in the *P. aeruginosa* population (Wiehlmann et al., 2007) by genome sequencing of 57 clone C and 42 clone PA14 isolates. The spatiotemporally unrelated strains were isolated during the last 30 years from the aquatic environment, acute infections or chronic airway infections in individuals with CF or chronic obstructive pulmonary disease (COPD) (Table S3).

Clonal conservation of the chromosomal frame

We calculated the length of syntenic segments with 100% sequence identity between pairs of either clone C, clone PA14 or clonally unrelated strains (Hilker et al., 2015). The completely sequenced NN2, PA14 and PAO1 genomes were taken as reference. Fig 2a shows the normalized frequency distribution of the length of fragments with identical sequence. Shared segments with absolute sequence identity, commonly called 'haplotypes' in case of diploid genomes, were short with a median size of 100 nucleotides if pairs of unrelated clones were compared (Hilker et al., 2015). In contrast, when either two clone C or two clone PA14 strains were compared, they shared large blocks of identical sequence with a median size of 99 kb (clone C) or 163 kb (clone PA14). Thus haplotypes are 1000-fold longer within a clonal complex than between unrelated clones. The chromosomal frame is conserved among members of a clonal complex, and in only a few cases the gene contig was disrupted by larger deletions (Fig 2b). Conversely, shared haplotypes between unrelated clonal complexes are smaller than the average gene length suggesting an unrestricted gene flow in the *P. aeruginosa* population by recombination.

Next, the haplotypes were used to visualize the relatedness of strains in a split-tree (Puigbò et al., 2012) (Fig. 3a, 3b). Tree topology is similar for the two clones. The majority of strains form a star-like structure of closely related independent singletons. Five (C) and six strains (PA14) are distant outliers. The preponderance of singletons suggests that most isolates of the clonal complex diverged from a common ancestor by few independent events.

Intraclonal diversity of core and accessory genome

The median intraclonal sequence diversity at the single nucleotide (SNP) level was found to be 0.372‰ (range 0.002 ‰ - 0.789 ‰) in clone C and 0.024 ‰ (range 0.008 ‰ - 0.973 ‰) in clone PA14 strains (Table 2). Most sequence diversity is caused by genomic islands (GI) and a few regions of genome plasticity (RGPs) of the accessory genome (Fig. 3c, 3d). If we only considered the core genome common to all *P. aeruginosa* the sequence diversity was calculated to be 8×10^{-6} for clone C and 2×10^{-5} for clone PA14 which is more than hundred-fold lower than the sequence diversity among unrelated clones (Hilker et al., 2015). In other words, the core genome is highly conserved in a clonal complex and differs by just a few dozen SNPs from one strain to another.

Intraclonal diversity was higher in the accessory genome not only in terms of sequence diversity, but also – and more importantly – in terms of the repertoire of GIs and RGPs of the individual strains. The composition of the accessory genome was nearly as variable among clone C or clone PA14 strains as between unrelated clones (Fig 4c). Figures 4a, 4b and supp. Figures 3 and 4 visualize the combinatorial composition of the accessory genome. The composition of the accessory genome segregated for 26 of the 100 isolates with spatiotemporal

origin, i.e. three pairs, five trios and one quintet were sampled each within a 3-year time period in a geographic area approximately equivalent to the size of Switzerland.

The highest sequence diversity was observed in the mobile integrative and conjugative elements (ICE) of the types pKLC102, PAGI-2 and PAPI-1 which co-exist as extrachromosomal elements and spread across species and genus barriers by horizontal transfer (Qiu et al., 2006; Klockgether et al., 2007; Pradervand et al., 2014). Consistent with their access to a large pool of proteobacterial hosts more SNPs were fixed in these mobile ICEs than in any other GI or RGP of the clone C and PA14 genomes. When we removed the SNPs located in these three ICEs from the data set, the SNP statistics of the two clones converged indicating that similar mechanisms of mutation and repair exist in C and PA14 (Table 2).

Intraclonal purifying selection of SNPs

Next, we examined the ratio d_S/d_N of synonymous to non-synonymous substitutions in the coding region of the 100 clone C and PA14 strains. The proportion of synonymous substitutions was higher in all strains than the expected value $d_S/d_N = 2.9/9.1$ of random mutation, but two groups could be clearly distinguished (Fig 3e). In the larger cluster of 70 strains the overrepresentation of synonymous substitutions was within the upper confidence interval of the expected value, but in all other strains including those with the largest intraclonal sequence diversity the portion of synonymous SNPs increased even further with the total number of SNPs. The logarithmic regression line also properly described the d_S/d_N ratio of clones other than C or PA14. These data give a hint on the evolution of coding genes in *P. aeruginosa.* When we compare intraclonal sequence diversity against a clonal reference genome, neutral substitutions are more likely to be fixed than amino acid substitutions. This trend of purifying selection against non-synonymous substitutions increases with the total number of SNPs until a d_S / d_N ratio of about four is reached that is typical for interclonal sequence diversity between unrelated clonal complexes.

The comparatively higher proportion of amino acid substitutions within as opposed to between clonal complexes also showed up in divergent frequencies of the type of change ($P < 10^{-6}$) and the functional category of the affected protein (Table 3, Fig 3f). This finding is plausible in the context of the divergent time scales of the evolution of the *P. aeruginosa* core genome and its clonal complexes. The few mostly clade-specific non-synonymous mutations in a clonal complex primarily affect proteins involved in bacterial communication with its environment which indicates their habitat-related emergence in a recent ancestor at the time scale of days to decades. Conversely, the retained protein variants of the core genome are the result of purifying selection over millions of years.

Intraclonal hot spots of mutation

Next, we searched for hotspots of mutations in the core genome of the 100 strains (Table S4). Besides numerous phage and plasmid-derived proteins found in both clones, the heavy metal ion efflux protein CusA, the cyclic-di-GMP phosphodiesterase BifA (Kuchma et al., 2007) and the key regulator of quorum sensing LasR (Williams et al., 2009) were identified in clone C as the targets for recurrent non-conservative amino acid substitutions some of which should modify structure and/or function. For example, in the case of LasR the amino acid exchanges include the helix-breaking incorporation of a proline which either affects conserved positions in the alpha helices 8 and 10 or are located in the binding pocket of the homoserine lactone autoinducer (Bottomley et al., 2007).

Alternatively to this search for genes with numerous amino acid changes, the genomes were scanned for segments with significantly elevated SNP rates (FDR < 0.05). The clone C strain panel carried twelve hot spots of mutations ten of which affecting intergenic regions or phage- and plasmid related genes (Table S5). The two genes in the core genome were the transcriptional regulator PA2020 and again *lasR.* Of the 60 segments found in the PA14 strains, 44 were located in the core and 16 in the accessory genome. Hot spots of mutations were predominantly phage- or plasmid-related genes and functionally to date uncharacterized open reading frames. Strain PT2 had accumulated almost all SNPs found in the 22 genes flanking RGP31 suggesting that the SNPs had been acquired from another clone by recombination. In contrast to this singular case, numerous strains of the PA14 clone harboured SNPs in the three hotspots of mutation of functionally characterized genes, i.e. *pchF, rocs2* and *pelA* indicating diversifying selection. Their gene products are involved in the communication of *P. aeruginosa* with its environment. PchF contributes to the non-ribosomal biosynthesis of the siderophore pyochelin (Patel et al., 2001), the transcriptional sensor RocS2 controls the biogenesis of CupC fimbriae and multidrug transport (Sivaneson et al., 2011) and PelA deacetylates the Pel exopolysaccharide which is essential for biofilm formation (Colvin et al., 2013).

Strain-specific gene repertoire

The strain-specific acquisition of genes could generate gain-of-function traits that modulate the fitness, lifestyle and metabolic competence of the clonal complex. And indeed, the provision of extra genes to the individual strain was found to be substantial in both the clone C and the clone PA14 complex (Table S6). An average PA14 or C strain had taken up 170 and 103 genes, respectively ($P < 0.001$ for the comparison PA14 vs. C). The majority of closest orthologs was identified in other *P. aeruginosa* clones or other members of the Pseudomonas genus (Fig. 4d). Phylogenetically more distant taxa contributed to the residual 20%.

Table 3 summarizes the total repertoire of strain-specific genes in the two clonal complexes sorted by functional category. Genes related to mobile genetic elements like phage or plasmids and hypotheticals of yet unknown function were significantly overrepresented among the strain-specific genes. This finding was expected because phages, transposons and plasmids are the

common vehicles to provide genes to an individual strain of a clonal complex by horizontal gene transfer. Conversely, genes that encode elements of transcription or of intermediary metabolism were rarely or not identified among the strain-specific genes indicating that the genetic repertoire of the core genome is essential and comprehensive to cope with the basic requisites of cell growth and metabolism of *P. aeruginosa* (Table 3). However, we noted a differential repertoire of genes promoting the metabolic competence of the bacteria to metabolize substrate. The clone PA14 strains harboured a larger number of genes involved in amino acid or fatty acid metabolism whereas the clone C strains had a larger genetic repertoire for carbohydrate metabolism (Table 3). It is textbook knowledge based on studies on a few reference strains like PAO1 that *P. aeruginosa* prefers amino acids and fatty acids as carbon source of intermediary and energy metabolism ('catabolite repression control') (Linares et al., 2010). Our data indicates that the repression of the uptake and catabolism of sugars may not apply to all *P. aeruginosa* and that some clonal lineages like the most common clone C may compensate the core genome-predetermined limitations in the utilization of sugars by the horizontal acquisition of genes of carbohydrate metabolism or by mutation of key regulators the latter having been reported for *P. aeruginosa* residing in CF lungs (Silo-Suh et al., 2005).

Intraclonal fitness

The preceding section describes intraclonal diversity in terms of the genomic make-up. Next we wanted to explore how intraclonal genome variation translates into differences in fitness among the members of a clonal complex. To accomplish this goal, sets of clone C or clone PA14 strains were grown together in planktonic culture.

To avoid the trivial outcome that fitness is governed by the individual proficiency of a strain to produce bactericidal pyocins, the weapon to kill other *P. aeruginosa,* and the corresponding neutralizing antitoxin (Michel-Briand et al., 2002), each strain was tested in its profile of production of and susceptibility to pyocins. All 57 clone C and 42 PA14 strains shared one clone-specific pyocin locus with each other and in addition the PA14 genomes differed in the presence or absence of two further pyocin operons. Unexpectedly the genotype-phenotype correlation was not stringent. The strain-specific production of pyocins and antitoxins was variable, and thus finally the largest subgroups of pyocin-tolerant strains consisted of just 10 clone PA14 and 32 clone C strains. These panels of strains each of which adjusted to the same optical density in the inoculum were grown for 48 h in either a mineral minimal medium or a nutrient-rich LB medium. The percentage of each strain in the sample at time points 0 and 48 h was determined by shotgun sequencing of the clone C or clone PA14 metagenomes and subsequent evaluation of the frequency distribution of single nucleotide sequence variants (Table S7). The *a posteriori* sequence analysis taught us that the surrogate parameter 'optical density' had significantly over- or underestimated the number of bacteria for three or eight strains, respectively.

Fig 5 depicts the fold change of the contribution of individual strains to the community after two and five days of co-culturing. The change of the relative abundance of individual strains covered a broad range from a 1,000-fold depletion to a more than tenfold enrichment in mineral medium as well as in LB broth in both the clone C and clone PA14 communities whereby for most strains their portion was not affected by the shift from exponential to stationary growth. Of the eight and eleven strains which were the most or least dominant strains under the various conditions, single clone C and PA14 isolates were consistently highly and two C and PA14 strain pairs consistently lowly abundant in both media. Winners and losers were examined for peculiar genomic features. The clone C winners were endowed with rare amino acid substitutions or frame-shifts in key regulators of lifestyle (LadS, CbrAB, several strains), extra weapons such as a lytic phage (110D4) or plasmid (MCF747) or extra copies of stringent response elements (81P29PA) (Table S6). The worst growing clone C strains exhibited non-conservative amino acid changes in enzymes involved in DNA supercoiling, LPS biosynthesis (WbpM, ArnA) or sensing of environmental cues (WspR). Of the four losers within the clone PA14 panel, two strains (31, 106120) shared a variant of the major genomic island PAPI-1, but otherwise each strain exhibited a broad spectrum of individual non-synonymous SNPs. The two clone PA14 winners were characterized by the highest similarity of its genome to the strain PA14 blueprint (CF1) or the largest accessory repertoire of transcriptional regulators, NADPH-dependent oxidoreductases and heavy metal ion efflux systems (39115), respectively. In summary, in accordance with the star-like structure of the dendrogram (Fig. 3a, 3b) the winners and losers did not share a group-specific repertoire of SNPs or genes, but rather carried individual genetic elements that governed their fitness to grow and to persist in the presence of other members of the clone. In particular, no association of fitness with habitat or geographic origin was observed.

Discussion

This first extensive study of intraclonal genome diversity of a cosmopolitan bacterium revealed a highly conserved core genome and a highly versatile accessory genome of its most common clonal complexes. *P. aeruginosa* is an ubiquitous microorganism that is equipped with broad nutritional capabilities, stress tolerance and an arsenal of virulence effectors (Ramos et al., 2004 – 2015). Members of the dominant clones C and PA14 have been isolated worldwide from soil and aquatic habitats and the animate surfaces of plants, animals and humans (Wiehlmann et al., 2007; Cramer et al., 2012). For our study, we selected isolates from the environment and from acute and chronic human infections. Unexpectedly the origin of the strain was not predictive of whether it was fitter than its clonal peers to grow and to persist in nutrient-rich or nutrient-poor planktonic cultures. Within the clonal community freshwater isolates had no collective advantage to outcompete clinical isolates in mineral medium, and conversely the isolates that were retrieved from chronically infected airways of subjects with COPD or cystic fibrosis were together not more proficient than the environmental strains at growing under

eutrophic conditions. This data demonstrates that it is the evolutionary history of the individual clade rather than the adaptation to the most recent habitat that determined the relative fitness of a strain within its clonal complex. Starting with approximately the same number of bacteria per strain in our competition, the fraction of strains ranged by up to 10,000-fold after just five days of planktonic co-culturing. Each strain was equipped with an individual set of genomic islands and RGPs which encode one hundred or more unique genes not shared with any other clone mate. The variable composition of the accessory genome thus strongly modified the fitness of the individual strain to compete with other clone members.

Lateral gene transfer turned out to be the driving force of intraclonal differentiation. On the contrary, the core genome with its sequence diversity of about 10^{-6} was virtually identical among members of the clone C or clone PA14 communities irrespective of their spatiotemporal origin. The few SNPs in the core genome were mostly strain-specific and the *de novo* coding variants were subject to purifying selection. In conclusion, the two dominant worldwide distributed *P. aeruginosa* clones are probably so successful at colonizing all aquatic habitats and mucosal surfaces on earth because their genome combines an almost invariant core with the flexible gain and loss of genetic elements that spread by horizontal transfer.

Experimental procedures

Strains. Specimens from human hosts were plated on blood, chocolate and MacConkey agar plates. An isolate was identified as *P. aeruginosa* on the basis of colony morphology, absence of lactose fermentation, presence of oxidase, growth at 42°C and the API20 NE system (BioMerieux, Nürtingen, Germany). *P. aeruginosa* was isolated from inanimate habitats by filtering of water samples (Selezska et al., 2012). Filters were placed on agar plates with Pseudomonas-selective medium (OXOID)(Pirnay et al., 2005) and incubated for 36 h at 37°C. Putative *P. aeruginosa* colonies were subcultured on the selective medium and typed by *P. aeruginosa* specific PCR (De Vos et al., 1997). Subcultures of *P. aeruginosa* isolates were stored at –80°C in LB broth supplemented with 17% (v/v) glycerol (Cramer et al., 2012) Strains were genotyped by a custom-made microarray (Wiehlmann et al., 2007) prior to genome sequencing.

Culture media. Bacteria were grown in either M9 medium (0.681 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 0.05 g/L NaCl, 0.1 g/L NH_4Cl, 0.1 g/L $MgSO_4$, 0.01 mM $FeSO_4$, 30 mM sodium succinate), liquid LB, tryptic soy broth or on solid LB agar plates.

Batch culture fermentation. For RNA-seq. experiments bacteria were grown in a BIOSTAT B reactor (Sartorius, Göttingen, Germany). *P. aeruginosa* NN2 was precultured with an initial optical density of 0.225 (550 nm) in 50 mL broth for 1 h at 37°C with shaking at 125 rpm. $1x10^7$ cfu thereof were inoculated into 1.5 L broth with 750 µL antifoam (Struktol). Bacteria were grown

at 37°C in batch culture at constant pH (7.0), agitation (400 rpm) and aeration (compressed air 0.3L/min). Growth was monitored offline by OD and CFU/mL and online by pO_2 and redox status. For RNA extraction, three samples were harvested in parallel at mid exponential and early stationary phase (determined by pO_2).

Assessment of pyocin production. Groups of clone C and PA14 strains were assembled based on the criterion of shared genomic pyocin loci. Each group member was grown overnight in LB broth at 37°C with shaking (150 rpm). The next morning 2.5 ml aliquots of each group member were mixed with fresh medium in one 100 ml flask and incubated for a further 24 h at 37°C and 150 rpm. Then the bacteria were precipitated by centrifugation and the supernatant was concentrated to half of its initial volume in a centrifugal evaporator. Agar plates with dried bacterial lawns of 100 µl overnight cultures of all single clone C or clone PA14 strains were inoculated with 40 µl drops of concentrated supernatants prepared from the various bacterial mixtures and dried. After overnight incubation at 37°C the plates were examined for halos indicating the inhibition of bacterial growth by the pyocin-containing supernatant. Thirty-three clone C and ten clone PA14 strains that showed no mutual inhibition of growth were selected for the competitive fitness experiments.

Growth competition experiments. LB agar plates were inoculated in parallel with loops of frozen glycerol stocks of different clone PA14 or clone C strains and incubated overnight at 37°C. Flasks with 5 ml LB broth were inoculated in parallel with a loop of bacterial lawn taken from one plate and incubated for the following 8 hours at 37°C with shaking. Aliquots of each clone PA14 or clone C strain were transferred into LB or M9 medium adjusted to a final OD_{578} = 0.1 and distributed into six technical replicates. Bacteria were grown aerobically in 25 ml flasks at 37°C with shaking (150 rpm). After 12, 24, 36 and 48 h the bacterial mixtures were inoculated into fresh media (final OD_{578} = 0.1). Samples were taken at time points 0, 48 and 120 h.

Preparation of DNA or RNA. Genomic DNA was isolated from *P. aeruginosa* according to a protocol optimized for Gram-negative bacteria (Ausubel et al., 1994). Total RNA (>18 nt) was extracted from 5 mL bacterial culture by using RNAprotect Bacteria Reagent and the RNeasy Midi Kit (Qiagen) according to manufacturer's instructions yielding high-quality RNA samples with a RNA integrity number of 7.6 or higher in the Bioanalyzer (Agilent, Bioanalyzer 2100 expert). Samples of 5 µg total RNA were treated with DNaseI to eliminate any residual DNA and then processed with the RiboZero Kit for Gram-negative bacteria (Metabion) to remove ribosomal RNA. RiboZero treated RNA was recovered by ethanol precipitation and the pellet was dissolved in 12 µL H_2O.

Sequencing. In case of the NN2 genome project, tagged paired-end and 3 kb mate pair libraries were prepared following the manufacturer's instructions and were sequenced on an Illumina Genome Analyzer II by GATC-Biotech (Constance, Germany). Furthermore single-end read libraries were prepared following the manufacturer's instructions and were sequenced on a

Roche 454 GS-FLX+ system. For RNAseq of strain NN2, strand-specific libraries were prepared from the RiboZero treated RNA samples with or without pretreatment with Terminator 5´-Phosphate-Dependent Exonuclease (TEX) following the 'TruSeq Stranded mRNA LT Sample Prep Kit' protocol (Illumina) and then sequenced on a HiSeq instrument.

All other genome and metagenome sequencing was executed in-house with a SOLiD5500 instrument. The manufacturer's standard protocol for fragment library generation was modified to overcome the substantial underrepresentation of sequences of a GC-content of 60% or more. One µg of bacterial DNA was sheared in a Covaris S2 system. End repair and size selection to an average of 200 bp fragment size were performed according to standard protocols (Fragment library generation, Life technologies (LT)/Thermo), but the next steps were modified. The dA tailing reaction was performed in ¼ of the standard volume with Stratec Taq Polymerase instead of the LT- dA tailing enzyme (DNA 9 µl; 5x Buffer (LT) 2.5µl, 10mM dATP 0.25 µl, Stratec Taq Polymerase 1.25 µl; 30 min; 68°C). The incubation conditions of the subsequent ligation were altered to increase life time and performance of the T4 ligase (dA-tailed reaction mix 13 µl; 5x Buffer (LT) 0.75 µl; each adaptor (LT, 1:20 diluted) 0.5 µl; 10mM dNTP 0.3 µl; T4 Ligase (NebNext, NEB) 0.8 µl; water 0.1µl; 12 h; 12°C; followed by nick translation (20 min; 72°C)). The generated fragment library was purified, amplified (5 cycles) and then bound to beads (EZBead System (LT); E120 scale, P2 post enrichment 17%) according to LT standard protocols. Sequencing was performed on a SOLiD 5500XL system (LT) with 75 bp read length and implemented ECC (Exact call chemistry (LT)). The produced sequencing reads were corrected using the SOLiD Accuracy Enhancer Tool (SAET).

Assembly of a single contig NN2 genome. The *de novo* assembly of the draft NN2 genome was started by first assembling the 454 sequencing data using the Newbler assembler (version 2.8) (Margulies et al., 2005). The minimum contig length was set to 150 bases and the internal read quality trimming of Newbler was used. The resulting scaffolds were combined and extended with the Illumina paired-end and 3kb mate pair data using the software SSPACE (Boetzer et al., 2011) with default parameters. The final scaffolds were aligned to the reference strain PAO1 using the contig arrangement software r2cat (Husemann et al., 2010).

Gaps with the predicted size of zero bp between two contigs were checked by performing a local alignment of all Illumina reads of this region with the short read aligner BWA (Li et al., 2009). Contigs with correct prediction and overlapping contigs were merged. The remaining gaps but two were closed by Sanger sequencing of PCR-amplified gene fragments (Goldstar polymerase (Eurogentec), fragments 250 – 400 bp, 5 cycles with elongation time 60 s). Of the final two gaps, one gap was closed by comparison with the known sequence of PAGI-2 (Larbig et al., 2002). The other gap was located in a region of multiple almost identical direct repeats. It was closed by recurrent searches for Illumina reads that linked the ends of the stepwise growing contigs.

Annotation of the NN2 genome. Automatic annotation of the closed contig was performed using the RAST-server (Aziz et al., 2008). The subsequent manual curation employed the Artemis tool (Rutherford et al., 2000). Gene coordinates were adapted to the gene length of orthologs present by March 2012 in the Pseudomonas Genome Database (Winsor et al., 2011) and the ascribed function was edited where applicable by screening of the biomedical literature listed in the PubMed database until December 2014. Other ORFs that lacked an ortholog in the Pseudomonas database or that were first identified by RNA-seq data were annotated by a matching BLAST result. The annotation file was deposited in the EMBL/EBI archive, accession no. PREBJ5222.

Transcriptional start sites (TSS) and non-coding RNAs (ncRNAs) were identified from the inspection of aligned cDNA reads with the genome viewer ReadXplorer (Hilker et al., 2014). The 5' UTR peak in TEX-treated libraries guided the mapping of TTS. ncRNAs were detected by significantly elevated counts of read contigs within and between annotated ORFs. A ncRNA was only counted if the same profile of cDNA reads was observed in technical and biological replicates.

Tetranucleotide usage in the NN2 genome. Tetranucleotide frequencies in the global NN2 genome and in a 300 bp window stretching from 250 bp upstream to 50 bp downstream of the start codon were counted with the CLC genomics workbench 7.0 **whereby in case of the latter only genes with a confirmed 5'UTR in the RNA-seq data set were considered. The observed counts were compared with the values if tetranucleotide frequency were only governed by mononucleotide content. The normalized difference between observed and predicted counts was visualized by color code for all 256 tetranucleotides sorted by** base stacking energy (Baldi et al., 2000).

Analysis of clone C and clone PA14 genomes.

Alignment of SOLiD reads. Untrimmed SOLiD reads were aligned to the strains NN2 or PA14 genomes by using the program NovoalignCS (www.novocraft.com) and the parameters -F CSFASTAnQV -r Random -c 16 -o SAM to create sam-files. Bam files were created from the sam files with picard-tools 1.68. Duplicates were removed using the samtools command rmdup.

Genomic islands (GIs) and regions of genome plasticity (RGPs). Single-end 75 bp SOLiD reads of clone C and PA14 isolates were aligned to a reference data set of known *P. aeruginosa* GIs and RGPs (Klockgether et al., 2011) using NovoalignCS. Uncovered reference regions were identified using the genomeCoverageBed utility from BEDTools (Quinlan et al., 2010). Coverage percentages of GIs and RGPs were calculated for each strain using an in-house Perl script. We noticed that the identified calculated sequence coverage was affected by the number of reads for each isolate. Therefore, the coverage percentages were normalized by dividing them by the probability *P* of covering a genome position with at least one read.

However, it was also desirable to determine which genomic islands and regions of genome plasticity are either absent of present in the selected isolates by either minimizing or amplifying coverage of partially covered features. For this purpose, inverse logit transformation was also applied to the coverage data. Genomic regions with values of 0.5 or greater were assigned as present. For pairwise comparison of strains, the absence or presence of a genomic region in both strains were scored with +1, whereas the presence of a genomic region in only one strain was scored with -1. Heatmaps were created using the gplots R package.

Identification of strain-specific genes. All reads that did not align to the corresponding reference strain or the known GIs or RGPs were separated with samtools view -b -f 4 and de-novo assembled with the program Velvet (Zerbino et al., 2008) with the k-mer sizes 27, 29, 31, 33, 35, 37, 39, 41. For each strain the best assembly result based on a large n50 number and a maximal contig length was chosen for further processing. False positive results were excluded by mapping the contigs against reference strain, GIs and RGPs with the long read mapping program CUSHAW2 (Liu et al., 2012). The remaining contigs were analyzed by a local blastx search against all bacterial genomes deposited in the UniProt database with the parameters -m 8 -e 1e-10 -a 8. To reduce the number of double entries, all hits with a sequence similarity of more than 90 % in at least 80 % of their sequence length were concatenated into one entry using an in-house script.

SNPs. SNPs were called in a three-step process using samtools (Li et al., 2009) from the bam-files. At first 'samtools mpileup' was performed with the options -g and -B (generate BCF output (genotype likelihoods); disable BAQ computation) followed by a filter step using 'bcftools view' with the options -c, -v, -g and -P (SNP calling (force -e); output potential variant sites only (force -c); call genotypes at variant sites (force -c); type of prior: full, cond2, flat [full]). The last step was the command `vcfutils.pl varFilter' to create the full SNP list of each strain without further parameters. From the full SNP-list all entries with a coverage of less than 4, a quality value below 50 and an allele frequency < 1 were removed. SNP-statistics were analyzed with the tool SnpEff 1.9.5 (Cingolani et al., 2009). Amino acid changes were grouped and analyzed according to the Dayhoff similarity index matrix, which describes the frequencies of amino acid substitutions between closely related proteins (Dayhoff, 1978).

Small insertions and deletions (indels). Indel calls were made on the basis of the results of the alignment of SOLiD read datasets to the PA14 and NN2 reference genomes. Variant calls were extracted from the existing sam-files using SAMtools and filtered for calls flagged as indels. All positions displaying a coverage of less than six at the respective position or showing the wildtype sequence in more than 50 % of reads were removed. The remaining candidate positions were then examined by manual inspection of the local alignment using the Integrative Genome Viewer (IGV) (Thorvaldsdóttir et al., 2013). Indel calls were considered as not reliable if one or more of the following criteria were observed: a) indel position too close (within length of one read) to an uncovered region; b) locus covered below six for the majority of isolates; c)

positions covered only by read ends; d) wildtype and indel reads found for all isolates; e) inconsistent position of the indel flagged by the reads; f) dominant calls of wild-type sequence erroneously neglected in output because of further sequence variations in cis. If indel calls were confirmed for several strains, the alignment was checked for the whole data set in order to trace the spread of the indel in the clonal complex or to recognize a unique sequence variant or sequencing error in the reference strain. Indels were subsequently annotated with SnpEff 1.9.5 to determine the affected genomic feature and potential frame-shifts in coding regions. Indels affecting protein coding genes were sorted by annotation class.

Large deletions. The genomes of the clone C and clone PA14 strains were scanned for deletions in 1,000 bp sliding windows using the script rpkmforgenes.py (Ramsköld et al., 2009). Hits were verified by manual inspection with the Artemis genome browser.

Length of syntenic fragments ('haplotypes'). Two matrices were constructed that contained columns of all quality-controlled clone C or clone PA14 SNPs ordered by genome position in the reference genomes and rows of the 58 clone C or 42 clone PA14 isolates, respectively. The value 0 was assigned to nucleotides that match with the reference and the value 1 was assigned to the nucleotide substitution. Next, all 1653 and 861 possible combinations of two clone C or clone PA14 genomes were compared in their similarity of SNP pattern. Haplotypes were identified by counting successive matches of binary pattern until the first mismatch. The number N of syntenic SNPs was then inserted at each SNP position of the haplotype. For this purpose a second matrix of haplotypes was constructed that consisted of all SNP positions as columns, the paired comparisons as rows and the numbers of syntenic SNPs, i.e. the haplotypes, as entries. This matrix was used to extract the haplotypes of the paired comparisons and to convert SNP synteny into physical length. Automatic analysis was performed with in-house Perl scripts.

Phylogenetic tree. SNPs with a map position in the NN2 or PA14 genomes were incorporated into the reference genome using the in-house script SequenceReplacer and concatenated to one file. The phylogenetic tree was created with the program Splitstree (Huson et al., 2006).

Structure of ncRNAs. Predicted secondary structure and thermodynamic stability of ncRNA variants were compared with the ViennaRNA websuite (Gruber et al., 2008).

Metagenome analysis. Competitive growth experiments were performed with pools of either clone C or clone PA14 strains. The percentage of the *n* individual strains in the sample was determined from SNP frequencies in the sequenced metagenomes. The data set allows to repetitively extract the composition of a clonal community of *n* strains by mass conservation law from 1. *n* linear equations of the cumulative frequency of strain-specific SNPs of the *n* strains; 2. $\binom{n}{2}$ linear equations of the cumulative frequency of SNPs shared by two of the *n* strains; 3. $\binom{n}{3}$ linear equations of the cumulative frequency of SNPs shared by trios; …. and so forth until *k.*

$\binom{n}{k}$ linear equations of the cumulative frequency of SNPs shared by *k* = half of the *n* strains. This set of $\sum_{n=1}^{k}\binom{n}{k}$ linear equations massively overdetermines the *n* unknowns, and consequently the abundance of strains in the pool can be calculated with high accuracy even though the coverage of reads at the individual genome positions of the SNPs may be poor. Since the phylogenetic tree of the clonal complexes showed a star-like structure, the practical evaluation focused on singletons and pairs as follows: The SOLiD reads were aligned to the NN2 and PA14 reference genomes using NovoalignCS. Pooled SNPs were identified using the GATK UnifiedGenotyper (McKenna et al., 2010) with the --sample_ploidy parameter set to the number of strains in the pooled samples.

Acknowledgements.

The authors would like to thank M. Griese (München), T.F. Murphy (Buffalo), J. Sikorski (Braunschweig) and C. Winstanley (Liverpool) for the provision of strains. This work was supported by grants from the Christiane Herzog Stiftung to N.C., the Deutsche Forschungsgemeinschaft (SFB 900, project A2) to B.T. and to S.S. and B.T. (SFB900, project Z1) and from the Bundesministerium für Bildung und Forschung (programme 'Medical Infection Genomics', 0315827A) to A.G. and B.T.. P.M.L. is a member of the graduate programme 'Infection biology' of Hannover Medical School.

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Contributions

S.F., N.C. and B.T. conceived the study. S.F., N.C., J.K., S.D., M.D., S.M., S.W. and L.W. performed experiments. S.F., P.M.L., P.C. and B.T. devised algorithms. S.F., P.M.L., P.C., C.F.D., A.G., R.H., T.S., O.T. and J.K. wrote scripts. S.F., N.C., P.M.L., P.C., S.D., C.F.D., R.H., P.C., J.K., and B.T. processed and evaluated the primary data. The manuscript was prepared by S.F. and B.T. with contributions by N.C., P.M.L., P.C., L.W. and J.K.. All authors read and approved the manuscript.

Competing financial interests.

The authors declare no competing financial interests.

References

American Thoracic Society. 2005. Infectious Diseases Society of America. Guidelines for themanagement of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. Am. J. Respir. Crit. Care Med. **171:** 388-416.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidmann, J.G., Smith, J.A., and Struhl, K. (eds). (1994) Current Protocols in Molecular Biology. New York, NY, USA: Wiley.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, RA., et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9**: 75.

Badrinarayanan, A., Lesterlin, C., Reyes-Lamothe, R., and Sherratt, D. (2012) The Escherichia coli SMC complex, MukBEF, shapes nucleoid organization independently of DNA replication. J. Bacteriol. **194**: 4669-4676.

Baldi, P., and Baisnée, P.F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. Bioinformatics **16**: 865-889.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**: 578-579.

Bottomley, M.J., Muraglia, E., Bazzo, R., and Carfì, A. (2007) Molecular insights into quorum sensing in the human pathogen Pseudomonas aeruginosa from the structure of the virulence regulator LasR bound to its autoinducer. J. Biol. Chem. **282**: 13592-13600.

Cingolani, P., Platts, A., Wange, le L., Coon, M., Nguyen, T., Wang, L., et al. (2009) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly **6**: 80-92.

Colvin, K.M., Alnabelseya, N., Baker, P., Whitney, J.C., Howell, P.L., and Parsek, M.R. (2013) PelA deacetylase activity is required for Pel polysaccharide synthesis in Pseudomonas aeruginosa. J. Bacteriol. **195**: 2329-2339.

Cramer, N., Klockgether, J., Wrasman, K., Schmidt, M., Davenport, C.F., and Tümmler, B. (2011) Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. Environ. Microbiol. **13**: 1690-704.

Cramer, N., Wiehlmann, L., Ciofu, O., Tamm, S., Høiby, N., and Tümmler, B. (2012) Molecular epidemiology of chronic *Pseudomonas aeruginosa* airway infections in cystic fibrosis. PLoS One **7**: e50731.

Dayhoff, M.O. (1978) Observed frequencies of amino acid replacements between closely related proteins. Atlas of Protein Sequence and Structure, 5, suppl. 3; National Biomedical Research Foundation, Washington D.C.

De Vos, D., Lim Jr, A., Pirnay, J.P., Struelens, M., Vandevelde, C., Duinslaeger, L., *et al*. (1997) Direct detection and identification of *Pseudomonas aeruginosa* in clinical samples such as skin biopsy specimens and expectorations by multiplex PCR based on two outer membrane lipoprotein genes, oprI and oprL. J. Clin. Microbiol. **35**: 1295-1299.

Gómez-Lozano, M., Marvig, R.L., Molin, S., and Long, K.S. (2012) Genome-wide identification of novel small RNAs in *Pseudomonas aeruginosa*. Environ. Microbiol. **14**: 2006-2016.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008) The Vienna RNA websuite. Nucleic Acids Res. **36**(Web Server issue): W70-74.

Hilker, R., Stadermann, K.B., Doppmeier, D., Kalinowski, J., Stoye, J., Straube, J., *et al*. (2014) ReadXplorer--visualization and analysis of mapped sequences. Bioinformatics **30**: 2247-2254.

Hilker, R., Munder, A., Klockgether, J., Losada, P.M., Chouvarine, P., Cramer, N., *et al*. (2015) Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. Environ. Microbiol. **17:** 29-46.

Husemann, P., and Stoye, J. (2010) r2cat: synteny plots and comparative assembly. Bioinformatics **26**: 570-571.

Huson, D.H., and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**: 254-267.

Klockgether, J., Reva, O., Larbig, K., and Tümmler, B. (2004) Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa* C. J. Bacteriol. **186**: 518-34.

Klockgether, J., Würdemann, D., Reva, O., Wiehlmann, L., and Tümmler, B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa.* J. Bacteriol. **189**: 2443-2459.

Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C.F., and Tümmler, B. (2011) *Pseudomonas aeruginosa* Genomic Structure and Diversity. Front. Microbiol. **2**: 150.

Kuchma, S.L., Brothers, K.M., Merritt, J.H., Liberati, N.T., Ausubel, F.M., and O'Toole, G.A. (2007) BifA, a cyclic-Di-GMP phosphodiesterase, inversely regulates biofilm formation and swarming motility by *Pseudomonas aeruginosa* PA14. J. Bacteriol. **189**: 8165-8178.

Larbig. K.D., Christmann, A., Johann, A., Klockgether, J., Hartsch, T., Merkl, R., *et al*. (2002) Gene islands integrated into tRNA(Gly) genes confer genome diversity on a *Pseudomonas aeruginosa* clone. J. Bacteriol. **184**: 6665-6680.

Lee, D.G., Urbach, J.M., Wu, G., Liberati, N.T., Feinbaum, R.L., Miyata, S., *et al.* (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome Biol. **7**: R90.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al*. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078-2079.

Linares, J.F., Moreno, R., Fajardo, A., Martínez-Solano, L., Escalante, R., Rojo, F., and Martínez, J.L. (2010) The global regulator Crc modulates metabolism, susceptibility to antibiotics and virulence in *Pseudomonas aeruginosa.* Environ. Microbiol. **12**: 3196-3212.

Liu, Y., and Schmidt, B. (2012) Long read alignment based on maximal exact match seeds. Bioinformatics **28**: i318-i324.

Margulies, M., Egholm, M., Altmann, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**: 376-380.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**: 1297-1303.

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

Patel, H.M., and Walsh, C.T. (2001) In vitro reconstitution of the *Pseudomonas aeruginosa* nonribosomal peptide synthesis of pyochelin: characterization of backbone tailoring thiazoline reductase and N-methyltransferase activities. Biochemistry **40**: 9023-9031.

Pirnay, J.P., Matthijs, S., Colak, H., Chablain, P., Bilocq, F., Van Eldere, J., *et al*. (2005) Global *Pseudomonas aeruginosa* biodiversity as reflected in a Belgian river. Environ. Microbiol. **7**: 969-980.

Pradervand, N., Sulser, S., Delavat, F., Miyazaki, R., Lamas, I., and van der Meer, J.R. (2014) An operon of three transcriptional regulators controls horizontal gene transfer of the integrative and conjugative element ICEclc in *Pseudomonas knackmussii* B13. PLoS Genet. **10**: e1004441.

Puigbò, P., Wolf, Y.I., and Koonin, E.V. (2012) Genome-wide comparative analysis of phylogenetic trees: the prokaryotic forest of life. Methods. Mol. Biol. **856**: 53-79.

Qiu, X., Gurkar, A.U., and Lory, S. (2006) Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. Proc. Natl. Acad. Sci. U.S.A. **103**: 19830-19835.

Quinlan, A.R., and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841-842.

Ramos, J.L. (ed.) (2004-2015) Pseudomonas. Vol. 1 – 7. Springer, Heidelberg – New York.

Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. **5**: e1000598.

Robinson, D.A., Falush, D., and Feil, E.J. (eds.). (2010) Bacterial populations genetics in infectious disease. John Wiley & Sons, Hoboken, New Jersey, USA.

Römling, U., Wingender, J., Müller, H., and Tümmler, B. (1994) A major *Pseudomonas aeruginosa* clone common to patients and aquatic habitats. Appl. Environ. Microbiol. **60**: 1734-1738.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrel, B. (2000) Artemis: sequence visualization and annotation. Bioinformatics **16**: 944-945.

Schmidt, K.D., Tümmler, B., and Römling, U. (1996) Comparative genome mapping of *Pseudomonas aeruginosa* PAO with *P. aeruginosa* C, which belongs to a major clone in cystic fibrosis patients and aquatic habitats. J. Bacteriol. **178**: 85-93.

Selezska, K., Kazmierczak, M., Müsken, M., Garbe, J., Schobert, M., Häussler, S., *et al*. (2012) *Pseudomonas aeruginosa* population structure revisited under environmental focus: impact of water quality and phage pressure. Environ. Microbiol. **14**: 1952-1967.

She, W., Mordukhova, E., Zhao, H., Petrushenko, Z.M., and Rybenkov, V.V. (2013) Mutational analysis of MukE reveals its role in focal subcellular localization of MukBEF. Mol. Microbiol. **87**: 539-552.

Silo-Suh, L., Suh, S.J., Phibbs, P.V., and Ohman, D.E. (2005) Adaptations of *Pseudomonas aeruginosa* to the cystic fibrosis lung environment can include deregulation of *zwf*, encoding glucose-6-phosphate dehydrogenase. J. Bacteriol. **187**: 7561-7568.

Sivaneson, M., Mikkelsen, H., Ventre, I., Bordi, C., and Filloux, A. (2011) Two-component regulatory systems in *Pseudomonas aeruginosa*: an intricate network mediating fimbrial and efflux pump gene expression. Mol. Microbiol. **79**: 1353-1366.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**.** Brief. Bioinform. **14**: 178-192.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G., *et al* (2007) Population structure of *Pseudomonas aeruginosa.* Proc. Natl. Acad. Sci. U.S.A. **104**: 8101-8106.

Williams, P., and Cámara, M. (2009) Quorum sensing and environmental adaptation in *Pseudomonas aeruginosa*: a tale of regulatory networks and multifunctional signal molecules. Curr. Opin. Microbiol. **12**: 182-191.

Winsor, G.L., Lam, D.K., Fleming, L., Lo, R., Whiteside, M.D., Yu, N.Y., *et al*. (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. Nucleic Acids Res. **39**(Database issue): D596-600.

Wurtzel, O., Yoder-Himes, D.R., Han, K., Dandekar, A.A., Edelheit, S., Greenberg, E.P., *et al*. (2012) The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. PLoS Pathog. **8**: e1002945.

Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**: 821-829.

Fig 1. The clone C NN2 genome. (**a**) Genome map. (**b**) Distribution of the length of the 5'-untranslated region in the strain NN2 (grey) and strain PA14 genomes. (black). (**c**) Comparison of the tetranucleotide composition in the whole NN2 genome (left) and in the segment of 250 bp upstream to 50 bp downstream of the start codon of all genes with an experimentally identified transcriptional start site.

Fig 2. Conservation of the core genome. (**a**) Normalized distribution of the length of 100% pairwise conserved sequence ('haplotype') in 58 clone C (n = 33,800 haplotypes), 42 clone PA14 (n = 9,510) and 20 clonally unrelated *P. aeruginosa* strains (n = 3,779,224). (**b**) Deletions in the core genome found in clone C (outer circle) and clone PA14 isolates (inner circle) from different habitats (full circle = chronic infection, open circle = acute infection, square = environment).

Fig 3. Single nucleotide sequence diversity of the clonal complexes C (reference: strain NN2 genome) and PA14 (reference: strain PA14 genome). (**a, b**) SNP-based phylogenetic trees of the clonal complexes C (**a**) and PA14 (**b**). (**c, d**) Cumulative Kaplan-Meier-plots of the SNP frequency along the genomes of clone C isolates (**c**) (n = 57) and of clone PA14 isolates (**d**) (n=40 and separately the two outliers PT2 and 158). Regions with pronounced sequence diversity are indicated by RGP or ORF numbers. SNPs in RGPs marked with an asterisk were not incorporated because of their large number of SNPs. (**e**) Comparison of intra- vs. interclonal sequence diversity: Plot of the ratio of synonymous to non-synonymous SNPs (d_s/d_n) vs. the total number of SNPs per strain. Clone C, clone PA14 and clonally unrelated strains (reference: strain PAO1 genome) are differentiated by symbol. The dotted line indicates the expectancy value of random mutation. (**f**) Normalized frequency of amino acid replacements within and between clonal complexes sorted by occupancy of Dayhoff similarity index.

Fig 4. Diversity of the accessory genome of the clonal complexes C and PA14. (**a, b**) The heatmap shows the presence (red) or absence (green) of genomic islands in (**a**) clone C and (**b**) clone PA14 isolates. Strains are arranged by the similarity of their repertoire applying hierarchical clustering with default parameters (R package). (**c**) Box-plot presentation of the similarity of the accessory genome within and between clonal complexes. For each strain a global score of relatedness was evaluated whereby the two strains were assessed of whether they were concordant (assigned value: + 1) or discordant (assigned value: - 1) for the presence or absence of each RGP or genomic island known from eight completely sequenced *P. aeruginosa* genomes. Please note the large overlap of scores of the intraclonal comparisons (C_C; PA14_PA14) with those of interclonal comparisons of 20 unrelated strains (20_20). (**d**) Origin of closest homologues of strain-specific genes.

Fig 5. Results of the fitness experiments. Fitness experiments in LB (a, b, upper panel) and mineral medium (MM, c, d, lower panel) of 32 clone C (b, d, right panel) and 10 clone PA14 strains (a, c, left panel). The figure displays the fold change of the contribution of individual

strains to the community after two and five days of co-culturing compared to the start of the experiment (open circle: isolate from acute human infection, closed circle: isolates from chronic infection, square: isolate from inanimate aquatic habitats).

Supporting Information

Supporting Text S1. The text describes the sequence diversity of ncRNAs and insertions, deletions and frame-shifts.

Supporting Figures:

Figure S1. **SNPs in ncRNAs.** (**a**) Kaplan-Meier-plots of the frequency distribution of ncRNA SNPs in the *P. aeruginosa* clone C and clone PA14 strain panels. (**b**) Histogram of the frequency of SNPs in individual ncRNAs in the *P. aeruginosa* clone C and clone PA14 strain panels.

Figure S2. Indel Frequency in the strain panel. Frequency of small indels in the clone C (left) and clone PA14 (right) strain panels differentiated by habitat and their localization within intergenic region, genes of annotation classes 1 or 2 and 3 or4, respectively.

Figure S3. Diversity of the RGPs of the accessory genome of the clone PA14 strains. The heatmap shows the presence (red) or absence (green) of RGPs in clone PA14 isolates. Strains are arranged by the similarity of their repertoire applying hierarchical clustering with default parameters (R package).

Figure S4. Diversity of the RGPs of the accessory genome of the clone C strains. The heatmap shows the presence (red) or absence (green) of RGPs in clone C isolates. Strains are arranged by the similarity of their repertoire applying hierarchical clustering with default parameters (R package).

Supporting Tables:

Table S1. Annotation of the NN2 genome. The table lists the annotated ORFs of the NN2 genome and yet undescribed ncRNAs. The table does not list ncRNAs first detected in *P. aeruginosa* PAO1 by RNAseq (Gómez-Lozano et al., 2012).

Table S2. 5'-untranslated regions of the NN2 genome. . Map positions of the transcriptional start site and the corresponding downstream coordinates of the next gene being either a singleton or the first gene of an operon in the *P. aeruginosa* NN2 genome. Only those transcriptional start sites are listed which were covered by at least 30 reads in TEX RNA-seq of NN2 bacteria grown in a fermenter with TSB medium (see Experimental Procedures).

Table S3. Origin and detailed SNP statistics of the investigated strains. The table lists the origin and isolation date of each strain and provides a detailed SNP statistics compared to the clonal reference (clone C: strain NN2, clone PA14: strain PA14).

Table S4. Amino acid exchanges of the strains. The table shows all amino acid exchanges found in the clone C and clone PA14 strain panel.

Table S5. Hotspots of mutation in the genomes. The table shows all genes with a high mutation frequency within the strain panels and their affiliation to core or accessory genome.

Table S6. Singular or shared genes that are absent in the reference genome and known RGPs. The table shows the additional genes of the strains. Each row lists (from left to right) the annotation and origin of the closest homologue, the strains harbouring the gene and up to ten more distant homologues, if applicable.

Table S7. Results of the fitness experiments. The percentage of individual strains in the samples was determined as follows: DNA extracted from the samples was randomly sequenced by high-throughput sequencing. The percentage of individual strains was calculated from the ratio of reads covering the strain-specific SNPs (Specific reads) to the total number of reads covering these genome positions (Number of total reads). In accordance with the star-like dendrogram (Figure 3) most SNPs only occurred as singletons or pairs allowing a straightforward quantitation. Only SNPs were considered that occurred in five strains or less.

Table S8. List of all SNPs in small RNAs. The table shows all SNPs located in ncRNAs detected in this work. The first base shows the reference sequence and the second one the SNP within the strain.

Table S9. ncRNA SNP statistics. The table shows number and frequency of SNPs in the ncRNAs of the strain panel.

Table S10. Stability of ncRNAs. The table shows in silico secondary structure and thermodynamic predictions of clone C and clone PA14 ncRNA SNP variants.

Table S11. Large deletions in the strain panel. The table shows the size and map positions of deletions in clone C and clone PA14 genomes including information about the deleted genes.

Table S12. Indels in the strain panel. The table shows for each strain the detected indels. Shared or strain specific indels are colored.

Figure 1

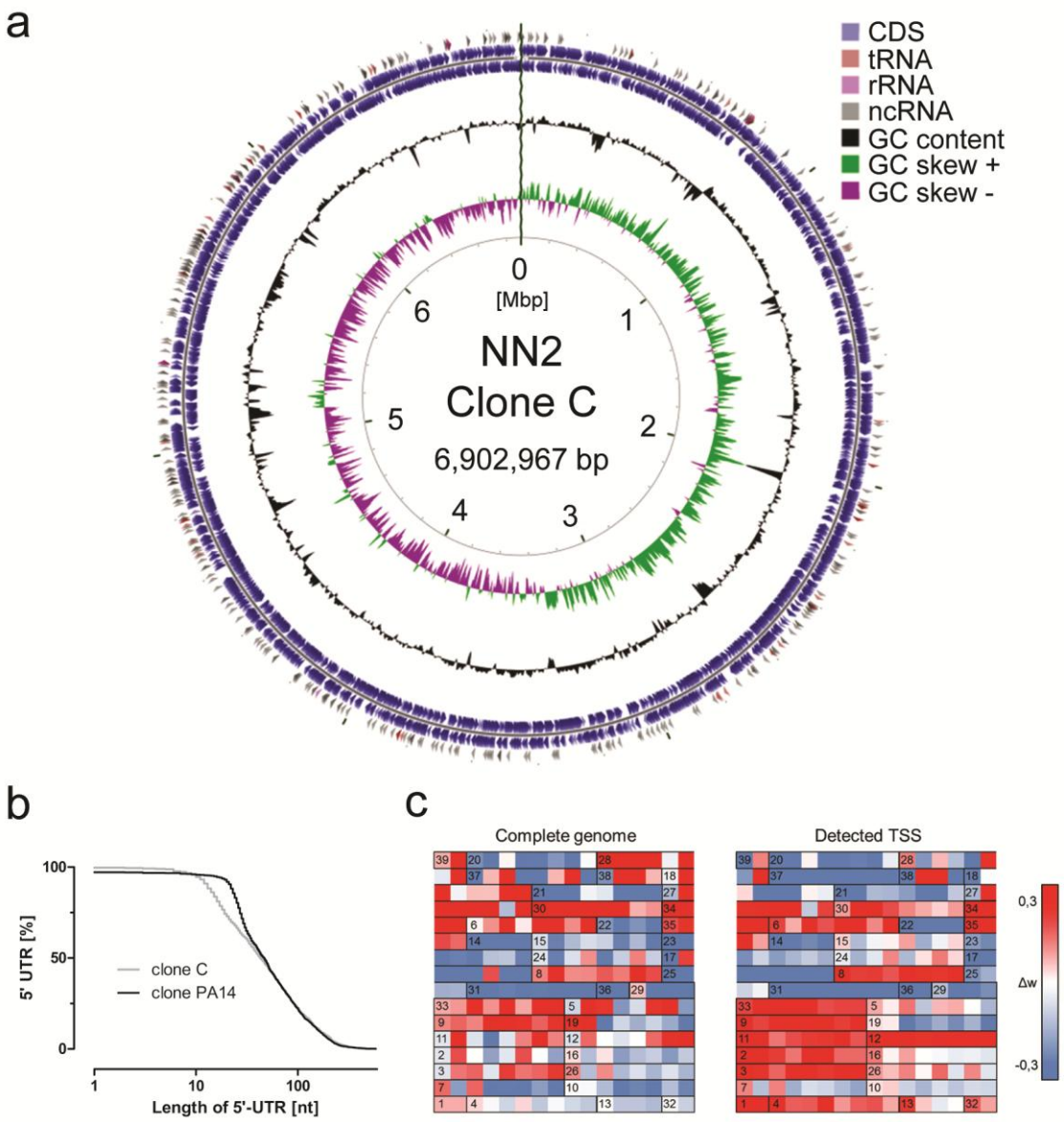


Figure2

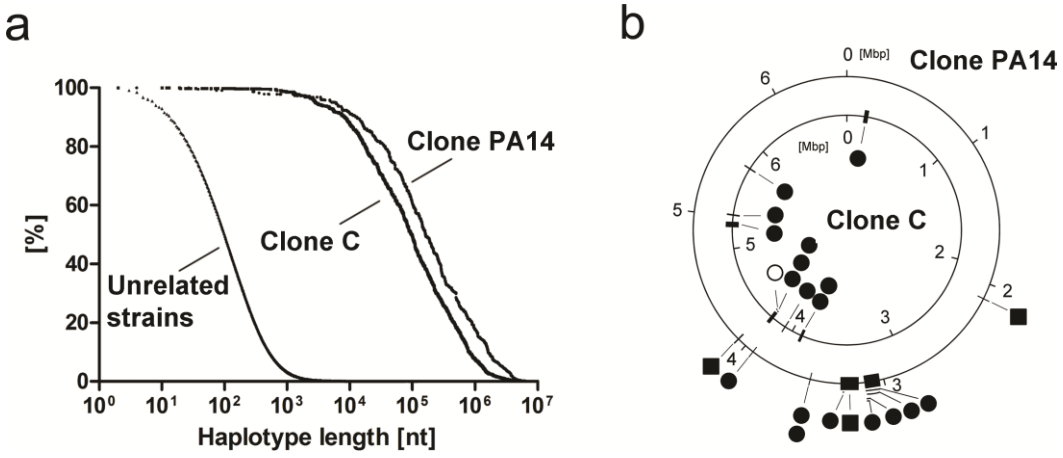


Figure 3

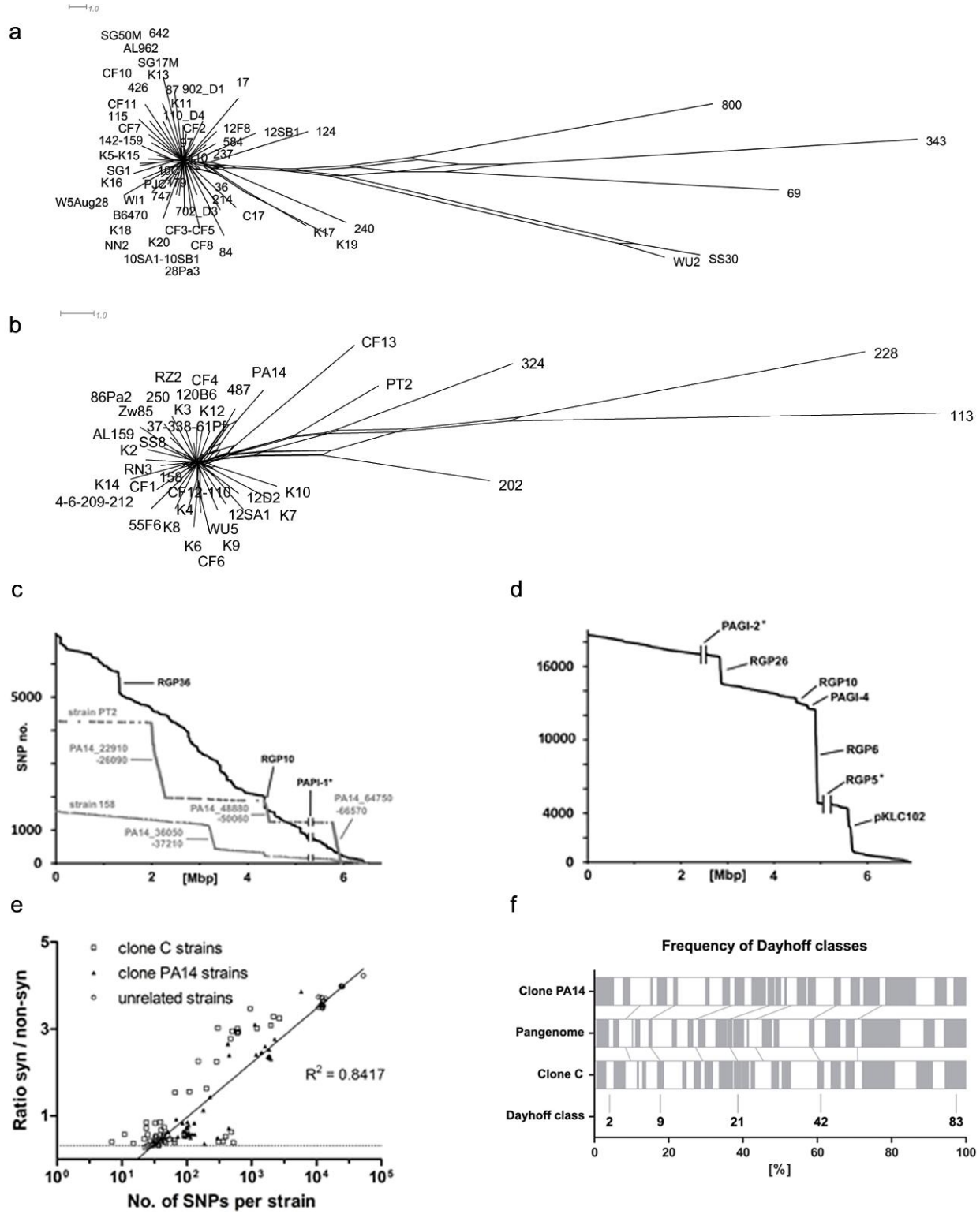

a
SG50M
642
AL962
SG17M
CF10
K13
426
902_D1
17
CF11
K11
115
110_D4
CF7
CF2
12F8
12SB1
124
142-159
584
K5-K15
237
SG1
K16
36
W5Aug28
WI1
747
214
B6470
702_D3
C17
K18
CF3-CF5
K17
240
NN2
K20
CF8
K19
84
10SA1-10SB1
28Pa3
800
343
69
WU2
SS30
b
CF13
324
228
RZ2
CF4
PA14
PT2
86Pa2
250
120B6
487
Zw85
K3
K12
113
AL159
37-338-61Pf
SS8
K2
RN3
K14
CF1
158
CF12-110
12D2
K10
4-6-209-212
K4
12SA1
K7
55F6
K8
WU5
K6
K9
CF6
202
c
RGP36
5000
strain PT2
SNP no.
PA14_22910
-26090
RGP10
PAPI-1*
PA14_64750
-66570
strain 158
PA14_48880
-50060
1000
PA14_36050
-37210
0
0
2
[Mbp]
4
6
d
PAGI-2*
16000
RGP26
RGP10
PAGI-4
10000
RGP6
RGP5*
4000
pKLC102
0
0
2
[Mbp]
4
6
e
5
clone C strains
clone PA14 strains
unrelated strains
Ratio syn / non-syn
3
R2 = 0.8417
1
10^0
10^1
10^2
10^3
10^4
10^5
No. of SNPs per strain
f
Frequency of Dayhoff classes
Clone PA14
Pangenome
Clone C
Dayhoff class
2
9
21
42
83
0
20
40
60
80
100
[%]

Figure 4

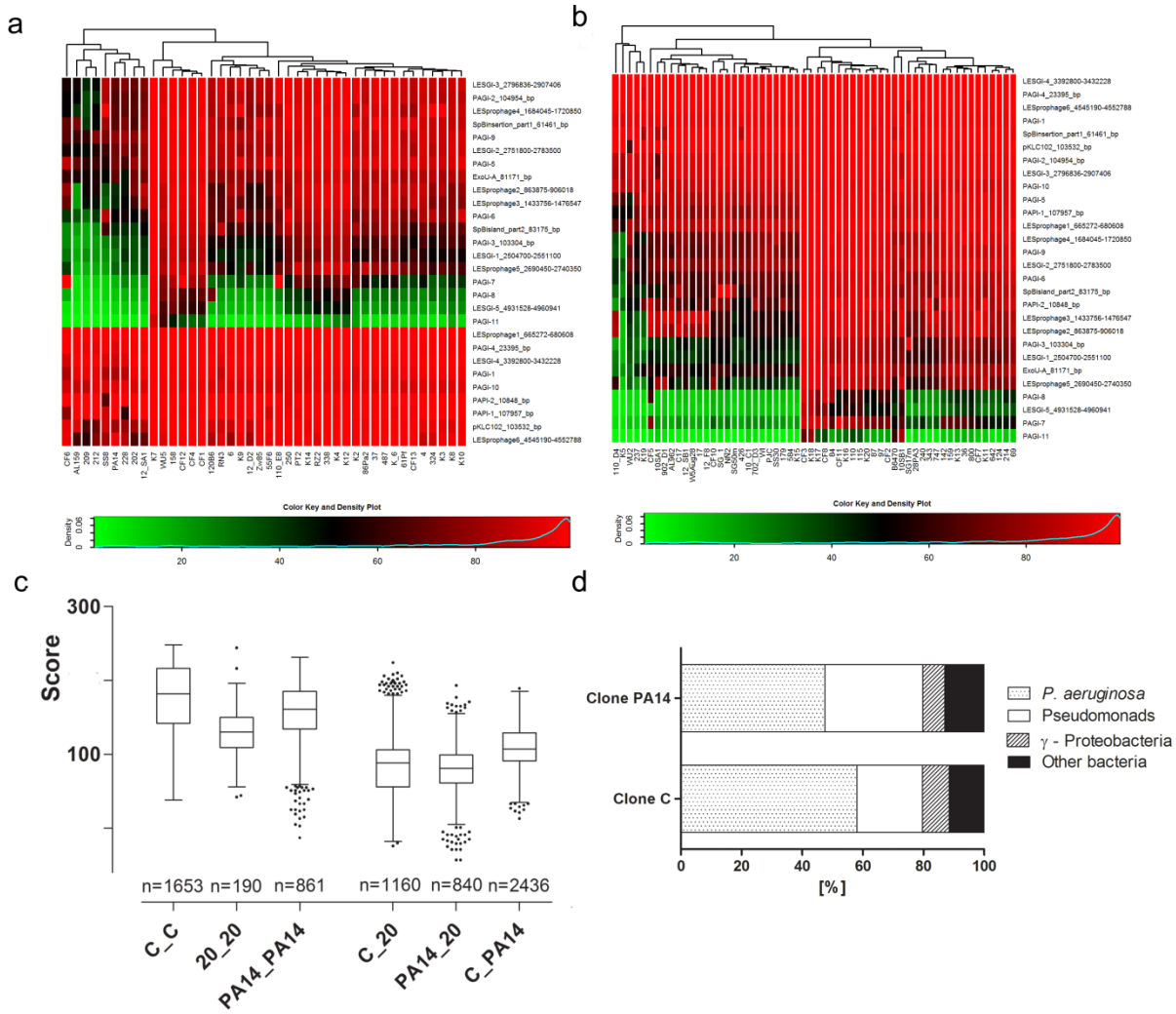
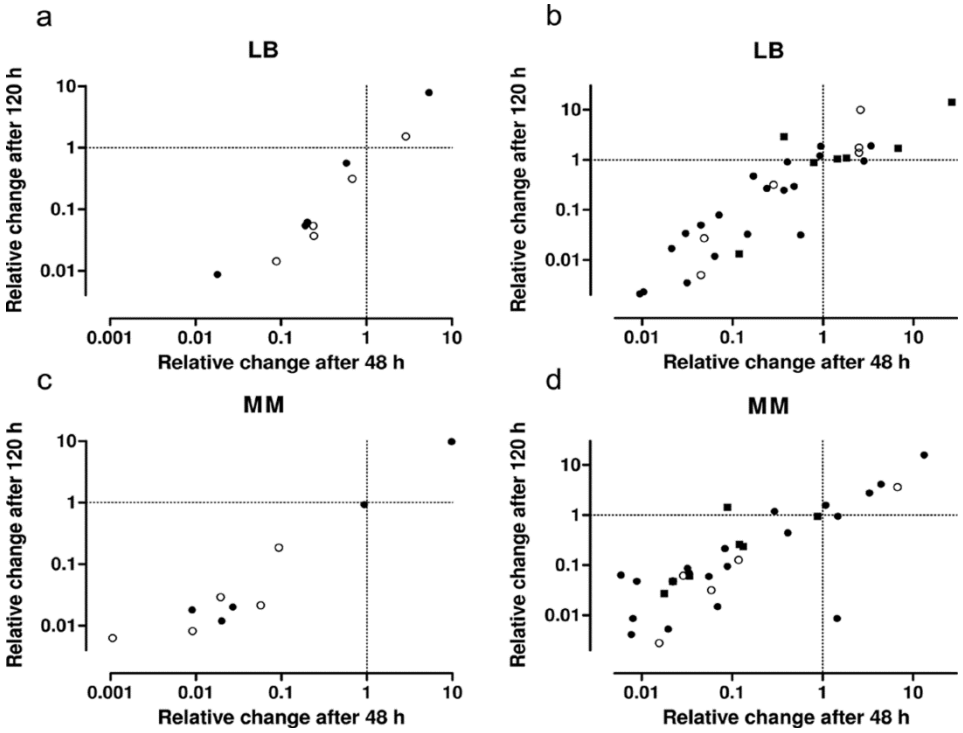


Figure 5



Chapter 6

Filtration and normalization of sequencing read data in whole-metagenome shotgun samples

6.1 Background

Next generation sequencing (NGS) has revolutionized and contributed significantly to expand our understanding of genomes. NGS has made affordable to sequence a whole metagenome sample and enable the identification and characterization of lowly abundant and unculturable bacteria or microbial communities within environmental samples. This technique is known as whole-metagenome shotgun (WMS) sequencing which has become an alternative to traditional 16S rDNA microbiome analysis.

WMS has significantly expanded our understanding of the diversity, composition and roles of the microbiome in different bacterial communities as well as in human health and diseases.

However, deep coverage variations and sequencing errors, such as GC content bias, can skew the relative abundance of bacterial diversity within the sample. GC content bias describes the proportion between fragments count and GC content found in sequencing data and it is considered not consistent between samples.

Several studies already showed the effect of GC content bias on fragment coverage in Illumina GA technology [88], but little is known about the effect of GC content in SOLiD technology. In case of Illumina, it has been observed that the fragment coverage increases with GC content, making regions with high AT content more difficult to sequence.

Another limitation using the WMS approach is the effect of horizontal genome transfer (HGT). HGT is an important phenomenon in prokaryotic evolution which enables the acquisition of new genes. Therefore, taxonomic assignment of sequences with horizontal gene transfer origin can generate overestimation of the bacterial abundances in specific loci of a single bacterial hit[89].

Finally, the normalization based on the lengths of the bacterial references genomes is also essential. In other words, large microbial genomes are more likely to be sequenced than small

genomes.

Thus, new computational challenges in the analysis of data become extremely needed and established to achieve quality controls and correct estimations of bacterial abundance using high-throughput sequencing.

6.2 About the paper

In this work we describe an efficient approach for the filtration and normalization of next-generation sequence data generated by SOLiD 5500xl technology, improving thereby accuracy in species identification and bacterial abundance estimation.

SOLiD sequencing produces short reads out of 75 nucleotides in length which are trimmed to a minimum length of 45 nucleotides and 40 bases with at least Q>=20. Sequences which do not satisfy the quality control step are filtered out.

Then our approach assigns each read to a specific reference genome, although, due to sequence similarity, reads can align to multiple genomes. In this case the reads were assigned to the lowest possible taxonomic level.

To study the GC content effect we created a GC bias model which use non-linear regression. The data used to create the model was collected from an empirical pooled sample with equal amounts of seven bacteria covering a broad spectrum of GC content (from 33% to 71 % GC content) and sequenced using the SOLiD 5500xl technology.

The model classified each read based on its GC content and the GC content of its hit (the bacterial reference genome to which it is assigned).

For filtration of clustered reads which could belong to genomic islands we present two complementary methods. The first method is based on a single-sample t-test of the mean distances between the reads mapped to the same reference genome. The second method uses a Poisson model to estimate the genome size of the reference based on the distances among all reads mapped to it. The difference between the actual and estimated genome sizes allows us to conclude whether the read mapping locations are spread distantly enough so that the origin from a genomic island shared by the reference genome and further yet unknown genomes can be excluded.

Finally, we perform a last step to normalize the data based on the genome length to which the sequence belongs to.

Author's contribution.

The study was conducted by P. Chouvarine I was ***mainly involved in the first steps of the pipeline, i.e. trimming of sequences and quality filtering, as well as the alignments of sequences.***

For Tables, please refer to the DVD attached to the thesis.

Filtration and normalization of sequencing read data in whole-metagenome shotgun samples

Philippe Chouvarine[1,2,*], Lutz Whiehlmann[1,2], Patricia Moran Losada[1] and Burkhard Tuemmler[1,2]

[1] Clinical Research Group, 'Molecular Pathology of Cystic Fibrosis and Pseudomonas Genomics', OE 6710, Hannover Medical School, Hannover D-30625, Germany.

[2] Biomedical Research in Endstage and Obstructive Lung Disease (BREATH), German Center for Lung Research, Hannover, Germany.

* To whom correspondence should be addressed. Tel: +49 511 532-7873; Fax: +49 511 532-6723; Email: chouvarine.philippe@mh-hannover.de

ABSTRACT

Ever-increasing affordability of next-generation sequencing makes whole-metagenome sequencing an attractive alternative to traditional 16S rDNA, RFLP, or culturing approaches to analysis of metagenomic samples. The advantage of whole-metagenome sequencing is that instead of analyzing a single 16S gene or other biomarkers it allows direct inference of the metabolic capacity and physiological features of the studied metagenome without reliance on the knowledge of genotypes and phenotypes of the members of the bacterial community. It also makes it possible to overcome problems of 16S rDNA sequencing, such as unknown copy number of the 16S gene and lack of sequence homology of the "universal" 16S primers to some of the target 16S genes. On the other hand, next-generation sequencing suffers from biases resulting in non-uniform coverage of the sequenced genomes, which, however, can be normalized. While there has been substantial research in normalization and filtration of read-count data in such techniques as RNA-seq or Chip-seq, to our knowledge, this has not been the case for the newly developing field of whole-metagenome shotgun sequencing. In this paper we present a model of GC-bias affecting sequencing reads in metagenomic samples and filtration and normalization techniques necessary for accurate quantification of microbial organisms in such samples.

INTRODUCTION

Metagenomics is the study of microbial communities in their natural habitat without isolation or cultivation of individual species (1). The boom of next-generation sequencing technologies makes it affordable to sequence with high coverage a whole metagenome of an environmental sample. This technique is known as whole-metagenome shotgun (WMS) sequencing. It is an attractive alternative to traditional 16S rDNA, RFLP, or culturing approaches to metagenomic analysis, because the techniques based on biomarkers rather than whole-genome analysis can suffer from inaccuracies due to copy number variation or lack of homology between the primers

and their targets. Moreover, these biomarker-based techniques are unable to assess the collective metabolic potential or the community population genetics, while it is possible using the WMS sequencing (2, 3). In addition, the WMS sequencing approach allows estimation of fungi and viruses in the sample, which is not possible with the biomarker-based metagenomic techniques.

The coverage of individual bacterial genomes comprising the metagenome will vary based on two factors: their abundance in the sample and sequencing artifacts, such as GC bias, fragmentation bias, the total amount of bacterial DNA sequenced, sequencing protocols, etc. Normalization of these biases can be used for correct estimation of bacterial abundance in the sample. Another pitfall in reporting bacterial abundance using the WMS approach is counting reads that are clustered only in a few loci of a single bacterial hit. Such reads are most likely located in genomic islands of horizontal transfer origin; therefore, such bacterial hits should be filtered out. Finally, lengths of the bacterial reference genomes also contribute to the likelihood of these genomes being sequenced, the same way that gene lengths affect the number of cDNA reads representing expression levels of these genes in an RNA-seq project (4).

In this paper we provide an overview of filtration and normalization procedures meant to improve accuracy in estimation of bacterial abundances in WMS samples. These methods are aimed to either discard or classify each read in a metagenomic sample to a correct species for the species-level analysis or to a strain if a strain-specific resolution is desired. Each classified read is given a weight based on its GC content and the GC content of its hit (the bacterial genome to which it is assigned). Finally, the length of the hit is applied to normalize the GC-weighted counts for accurate abundance estimation.

In the Filtration Methods section, we present two complementary methods for filtration of clustered reads potentially mapped to genomic islands. The first method is based on a single-sample t-test of the mean distances between the reads mapped to the same reference genome. The second method uses the Poisson model to estimate the genome size of the reference based on the distances among all reads mapped to it. The difference between the actual and estimated genome sizes allows us to conclude whether the read mapping locations are spread enough in order not to come from genomic islands incorporated into this reference genome.

In the GC Normalization section, we present a GC bias model that was created using non-linear regression from the empirical data collected by sequencing a pooled sample with equal amounts of seven bacteria with various GC contents using the SOLiD 5500xl technology. Similar models should be created for each sequencing platform as the GC biases are expected to vary for each of them (5).

In the Other Considerations section, we compare species level- and strain level WMS approaches. Namely, we discuss dealing with reads mapping equally well to multiple locations

in the same or multiple genomes at the species level analysis. Bacterial load assessment and genome length normalization are also discussed.

MATERIALS AND METHODS

DNA library preparation

Seven bacterial reference strains with different GC contents were obtained from the American Tissue Culture Collection or the in-house collection: *Burkholderia cepacia* (67% GC, ATCC 25416), *Escherichia coli* (50% GC, ATCC 25922), *Klebsiella pneumoniae* (57% GC, ATCC 10031), *Nocardia farcinia* (71% GC, MHH 442780), *Pandorea apitsta* (64.9% GC, RB-44), *Staphylococcus aureus* (33% GC, ATCC 25923), *Streptococcus pneumonia* (40% GC, ATCC 49619). Bacteria were grown until exponential phase in LB broth. DNA was isolated from the bacteria with the DNeasy kit (QIAGEN) following the instructions of the manufacturer. Yield of double-stranded DNA was quantified with the Qubit spectrofluorimeter (Invitrogen). Aliquots of five ng of each bacterial DNA preparation were added to 315 ng human DNA in a total volume of 130 µl low TE-buffer (Life Technologies).

Induced sputum was collected from subjects with cystic fibrosis during inhalation with aqueous hypertonic saline (6% v/v NaCl). The sputum sample was diluted 1:4 with phosphate-buffered saline/2% (v/v) mercaptoethanol at 4°C and incubated under shaking for 2 h on ice. The specimen was centrifuged at 3,800 g for 15 min at 10°C. After removal of the supernatant, the pellet was dried and then dissolved in 10 ml distilled water for 15 min at 4°C. The suspension was again centrifuged (3,800 g, 15 min, 10°C), the precipitate was dissolved in distilled water for 15 min at 4°C, pelleted and the pellet was transferred into an Eppendorf tube for incubation with DNase I (0.42 mL H_2O + 50 µL RD buffer (QIAGEN) + 35 µL DNase I) at 30°C for 90 min under shaking. The suspension was added to 10 ml SE-buffer and washed three times with 10 mL SE each by precipitation (3,800 g, 15 min, 10°C). The pellet was dissolved in 0.5 mL SE in an Eppendorf tube and precipitated again (12,000 g, 10 min, 10°C). DNA was extracted from this pellet with the Nucleo Spin Tissue Kit (Macherey & Nagel) by following the hard-to-lyse-bacteria protocol and stored at 4°C in TE buffer at 4°C until use.

Preparation of fragment libraries and sequencing were performed at the E120 scale according to the protocols provided by Thermo Life Technologies for SOLiD5500 instruments (generation of libraries: https://tools.lifetechnologies.com/content/sfs/manuals/4460960_5500_FragLibraryPrep_UG.pdf; emulsion PCR: Emulsifier, Amplifier http://tools.lifetechnologies.com/content/sfs/manuals/cms_102275.pdf; Enricher http://tools.lifetechnologies.com/content/sfs/manuals/cms_089261.pdf).

Analysis Dataflow

The SOLiD reads of metagenomic samples are first trimmed to variable lengths (no shorter than 45 bp) to have at least 40 bases with Q>=20. The trimmed reads are checked for contamination by Homo sapiens DNA by aligning them against the "1000 Genomes" Homo sapiens reference, which includes contigs unassigned to chromosomes. The unaligned reads are corrected using SOLiD's SAET utility, which increases the number of mapped reads by 40 - 50% in genomes of size 1Kbp - 200Mbp with coverage 10-4000x and read length 25 - 75bp (according to the manufacturer). The corrected reads are aligned against available reference genomes of bacteria, viruses, fungi, and known contaminates using the Novoalign (http://www.novocraft.com/) short read aligner. We set the –r parameter of Novoalign either to All (for the species level analysis) or to None (for the strain level analysis). This parameter determines the multiread strategy as described in the Other Considerations section below. Reads aligned to the bacterial references are filtered out if they do not pass the clustering tests described in the next section. The reads are normalized to correct the GC bias and reported as weighted counts per Mb of reference. The unaligned reads can still belong to species without reference genomes. These reads can be used for functional analysis after contig assembly with the subsequent blastx Uniprot search.

Filtration Methods

Bacterial genomes are known to actively recombine and incorporate genomic islands from bacteria of other strains or species. This can confound correct identification of species in a metagenomic sample. To avoid false attribution of mapped reads to references potentially containing genomic islands, it is necessary to filter out such bacterial hits that only have a few clusters or mapped reads. The following three steps were applied successively to achieve this:

(1) Single-sample t-test of the read start differences of neighboring reads. Ideally, all reads mapped to a genome should be uniformly distributed across its length. While some sequencing biases, such as GC bias or DNA fragmentation bias during the library preparation can distort a perfect uniform distribution of the read positions, such distortions are still negligible compared to the location bias of the reads mapping exclusively to the genomic islands. We can formally test it by calculating distances between the read start positions Δ of the neighboring reads and performing a single-sample t-test of the difference of the actual mean of such distances $\overline{\Delta}$ and the null hypothesis mean $\overline{\Delta}_{H0} = G/N$, where *G* is the genome length and *N* is the number of mapped reads.

From our experience setting the cutoff p-value to $p < 0.01$ removes most hits with the genomic island pattern of read clustering. However, some cases with a high number of islands can still produce high enough p-value to evade this test.

(2) Estimation of bacterial genome size based on the Poisson model. In this step we consider distances among all reads in a circular bacterial genome. In this case they are calculated as

$$\Delta = \begin{cases} p_2 - p_1, if|\Delta| < \frac{G}{2} \\ G - p_2 + p_1, otherwise \end{cases}$$

We can estimate the mean of distances between each possible pair of reads as

$$\overline{\Delta} = \frac{\sum_{1}^{N} \Delta}{\binom{N}{2}}$$

According to (6) this mean value can be used to estimate the size of a circular bacterial genome as

$$G_{pred} = \frac{4N\overline{\Delta}}{N \pm z\sqrt{N}},$$

where *N* is the number of mapped reads and z is the z-score, which can be set to 1.96 for a 95% confidence interval.

From experience we can see that when this confidence interval is used we can filter out hits with the actual genome size 10 times greater than the upper bound of the confidence interval of the estimated genome size $G_{pred\ \max}$. Such significant difference can be explained by read clustering within genomic islands, hence, such bacterial hits can be safely discarded. When $G_{pred\max}$ is 10 to 50% of the actual genome size, it is hard to programmatically determine if the reads are clustered in genomic islands. However, the heuristic described in the last filtration step can be applied before manual inspection of the alignment statistics.

(3) Filtration of hits based on the distribution of Δ counts of neighboring reads. This step is a heuristic that was formulated based on manual inspection of the uncertain cases from step 2. For hits with a small number of mapped reads (200 or less), if the distances between neighboring reads Δ that are shorter than 10 bp constitute at least 75% of distances Δ of any length, then the reads in this hit are mapped to genomic islands and the hit can be discarded.

GC Normalization

GC bias is the most significant bias adversely affecting coverage of GC-rich regions of a sequenced genome. Generally, G and C bonds are more stable than A and T, due to the fact that they have one extra hydrogen bond and their stacking interactions are quite different. Particularly, G-C pairing does not affect DNA duplex, while A-C pairing is always destabilizing

(7). It is present to various degrees in short-read next-generation sequencing technologies (5). Obviously, GC bias affects abundance estimates of bacterial genomes in metagenomic samples, particularly the ones with high GC content. As noted in (8) GC of the full PCR-amplified fragment rather than the forward read (or forward and reverse reads for paired-end sequencing) of this fragment primarily determines the GC bias. This also confirms previous findings of Aird *et al.* (9) that the PCR component of the GC bias is the one that contributes most, while the downstream instrument related bias is also present, but to a lesser degree. Moreover, there exists a global source of GC bias on the scale larger than the fragment length due to association with higher order structures of the DNA (8). We take into account this global source of GC bias and the PCR-induced, fragment GC bias by considering GC content of the genome to which the read is mapped, which is important in metagenomic samples with multiple bacterial genomes of a wide GC range. We take into account the post-PCR instrument GC bias by considering GC content of each read. In other words, a GC-rich read from a GC-poor locus typical of GC-poor genomes is more likely to be sequenced than a read with the same GC content, but located in a GC-rich locus. We confirmed this idea by pooling and sequencing equal amounts of seven bacteria ranging in GC content from 32.8 to 70.8% and calculating a GC bias curve for each of them (Fig. 2). Using the pooled sample rather than sequencing individual bacteria was important to simulate under- or over representation of bacteria in a metagenomic sample based on their GC content. The curves in Figure 2 were created by calculating the normalized coverage (GC bias) at each GC percentage point *i* using the CollectGcBiasMetrics utility from Picard Tools (http://broadinstitute.github.io/picard/). This program calculates normalized coverage for the case of sequencing a single genome as follows:

$$NormCov_i = \frac{Rst_i}{W_i} \Big/ \frac{Rst_{Total}}{W_{Total}},$$

where Rst_i is the count of read starts within windows of GC% *i* and W_i is the count of windows of GC% *i.* The ratio with the total values (Rst_{Total} / W_{Total}) normalizes the estimation to the average number of reads per window across the whole genome. To calculate normalized coverage of a single bacterium in a pooled sample we modified this formula as follows:

$$NormCov_{ij} = \frac{Rst_{ij}}{W_{ij}} \Big/ \frac{Rst_{Total} / m}{W_{Totalj}},$$

where *j* = 1,..., *m* are indices for *m* bacteria in a pooled sample. In this formula we distribute the total number of read starts in a sample Rst_{Total} equally among all bacteria in the pool instead of using the actual number of mapped reads to bacterium *j*, thus, accounting for potential under- or over representation of bacteria due to their GC content. However, the total number of windows

of various GC content is unique to a particular genome, therefore, W_{Totaj} is used in the formula. The read length of our sequences was 75 bp, though some reads were trimmed during the alignment. We used the default window size for short reads, which is 100 bp.

As shown in Figure 2, reads from GC-poor genomes were overrepresented, while the general bell-shaped-like curve of GC bias can be, at least partially, observed in all genomes. Our GC normalization model was designed by non-linear multiple regression of the experimental normalized coverage. In this model the dependent variable is the normalized coverage coefficient of the read (as defined by the equation above), while the GC content of the read and the GC content of the genome to which this read mapped are the independent variables. To normalize for GC bias each read should be divided by its normalized coverage coefficient.

The following formula was used for the regression:

$$NormCov(GC_R, GC_G; \theta) = B(GC_R; \theta_1, \theta_2, \theta_3, \theta_4) + \theta_5 GC_R + \theta_6 GC_R^{\,2} + \theta_7 GC_R^{\,3} + \theta_8 \log(GC_G),$$

where GC_R is GC content of the read, GC_G is GC content of the genome to which the read is mapped, $\theta_{1,\ldots,8}$ are regression coefficients, and *B* is the bell-shaped curve function of read GC content, defined as

$$B(GC_R; \theta_1, \theta_2, \theta_3, \theta_4) = \theta_1 e^{-0.5\left(\frac{GC_R - \theta_2}{\theta_3}\right)^2} + \theta_4.$$

The third degree polynomial in the regression formula accounts for imperfections in the bell shapes of the observed normalized coverage curves. The last term of the formula approximates the observed influence of genome GC content. Even though the *Nocardia farcinica* genome with the highest GC content (70.8%) appears to have higher normalized coverage values than two genomes with lower GC content (*Pandoraea apista,* GC 64.9% and *Burkholderia cepacia*, GC 67.7%), we suspect that these higher normalized coverage values are outliers, because genomes with higher GC content should be more difficult to sequence. Therefore, we used the (upside-down) logarithm function to approximate this dependency. The constructed model is only an approximation, therefore, to avoid extremely low predicted values, which would set unreasonably high weights to some reads, we set all predicted values less than 0.1 to 0.1.

The regression was performed using the nls function in R. The contour plot of the produced model is shown in Figure 3.

Other Considerations

Metagenomic analysis using WMS sequencing can be performed on the strain level or species level. To perform the strain level analysis only reads that are mapping uniquely to the strain-level references should be considered, thus achieving the desired specificity. However, from our experience, depending on the promiscuity of bacteria present in a metagenomic sample the percentage of multireads that map to multiple locations in the same or multiple genomes can vary from 5 to 96% of the total number of reads. Therefore, if the species level of analysis is desired, the multireads can be used for abundance estimates as long as each of them is counted only once and assigned to a single species. Our in-house Perl script performs such assignment by discarding multireads that map to more than one species and collapsing hits to multiple strains of the same species produced by a single read.

Another important consideration for accurate reporting of bacterial abundances in metagenomic species is the genome length normalization. This procedure is common in other analyses involving counting of short reads mapped to genomic features, e.g., RNA-seq, where RPKM (4) values are used for absolute levels of gene expression. Following the same logic, longer bacterial genomes will have a higher chance to produce sequencing reads. To account for this, the final bacterial abundances can be reported as GC-weighted read counts per Mb of reference.

Finally, estimation of bacterial loads in a metagenomic sample may be desirable, e.g., to assess an infection in an animate habitat. In this case, we can utilize the host background DNA to do this assessment and report absolute bacterial abundances per DNA content of a single host cell. If the host is human, the absolute abundance of each species per human cell can be calculated as follows:

$$A = \frac{6191.39 \ C_{GCperMbR}}{C_H},$$

where $C_{GCperMbR}$ is the GC weighted read count per Mb reference of the given species, C_H is the human read count, and 6191.39 is the length of a diploid human genome divided by a million (to account for the bacterial count scale).

RESULTS

Filtration of hits with clustered reads

We tested our three-step approach to filtering hits with reads potentially clustering in genomic islands by collecting percentages of filtered out reads for 30 cystic fibrosis sputum samples. If our method was filtering hits (and the associated reads) in samples with fewer reads more aggressively, this would indicate that our approach is biased by the sample read count and is

not applicable. However, we only found weak correlation between the number of mapped reads and the proportion of filtered out reads (Pearson's *R* = -0.237). The distribution of percentages of the filtered out reads vs. the number of mapped reads is shown in Figure 1.

GC normalization

To identify the effect of the GC bias on abundance estimates of a collection of genomes in a metagenomic sample we sequenced a pooled sample with equal amounts of DNA of seven bacterial genomes. Figure 2 shows superimposed curves of normalized coverage vs. GC percentage for each of the genomes. All data points with p-values less than 5% were removed.

As described in Materials and Methods, the normalized coverage reflects the GC bias of the genome at locations stratified by each GC percentage point. While the curves vary in shape, they all follow the same unimodal bell-curve pattern with various degrees of distortion. Notably, the *E. Coli* curve (genome GC 50.5%) remains relatively high in the GC range from 25 to 50%. This can be explained by the fact that sequencing kits are often optimized for human genome sequencing and it is known that GC content of 100 Kb fragments of the human genome can range from 35 to 60% (10).

Another clear observation from Figure 2 is that there is an inverse relationship between normalized coverage and genome GC content. Therefore, our GC bias model was designed with two independent variables: read GC content and GC content of the genome to which the read was assigned. The resulting nonlinear multiple regression model is described in detail in Materials and Methods. For our normalized coverage data, presented in Figure 2, the residual standard error was 0.3283 on 334 degrees of freedom. Conversion was achieved after 12 iterations with the conversion tolerance of 9.335e-06. The contour plot of the model is shown in Figure 3. Visual inspection of the contour plot shows that the model approximates the empirical data well and without undue over-fitting. As mentioned earlier, all values of the dependent variable approximated below 0.1 are programmatically set to 0.1.

The effects of GC- and genome length normalization are shown in Table 1 where the species abundances of a sample are reported as raw read counts, GC-weighted read counts, and GC-weighted read counts per Mb of reference. The three reported counts vary significantly for some species changing their rank number after the normalization steps.

Bacterial load estimates

For some metagenomic studies it is essential to assess absolute abundances of bacterial population. For example, examining the amount of bacteria found in lower airways of a cystic fibrosis patient can help diagnose the disease progression. Such assessment can be made by identifying the percentage of human DNA in the sample. Figure 4 shows relative and absolute abundances of bacteria in sputa taken from two patients with cystic fibrosis at 3-month intervals.

Please note that the presentation of relative percentages as it is common for most microbiome studies may lead to an erroneous interpretation.

DISCUSSION

Estimation of bacterial abundances by whole-metagenome shotgun (WMS) sequencing is based on the counts of reads mapped to a collection of entire bacterial genome references rather than a region of the 16S gene or other biomarkers uniquely identifying a genome. This poses a challenge since some reads mapped to one of the references can belong exclusively to genomic islands horizontally transferred from other organisms. Other challenges include uneven coverage of the reference genomes due to GC bias inherent to PCR and short-read next-generation sequencing platforms. Even varying lengths of the reference genomes affect likelihood of a read mapping to a particular bacterial genome reference, which is not a problem for 16S analysis where the sequenced variable regions of the 16S gene are all of the same length. However, we have shown that all these obstacles can be addressed by filtration and normalization procedures, thus leading to more accurate estimation of bacterial abundances in a metagenomic sample.

The structure of bacterial genomes is often dependent on frequent recombination due to significant evolutionary pressure to survive in hostile environments. Some bacterial genomes are particularly promiscuous in accepting horizontally transferred genomic islands, e.g., Pseudomonas or Burkholderia. We have shown that it is possible to identify false positive bacterial hits with reads clustered in their genomic islands. Application of our three-step filtration procedure removed 1.7 to 15.9% of such bacterial hits in the 30 cystic fibrosis airway samples (Fig.1).

Typical methods for GC correction rely on applying local regression (8, 11–13), e.g., LOESS, or quantile normalization methods (13, 14) to the data points created by read counts mapped to genes or non-overlapping windows of the reference sequence grouped into GC-stratified bins vs. the GC content of the bins. The raw read counts are then normalized, e.g., by calculating the correction value for each feature as the difference between the fitted value and the median across all bins. This strategy works well when a single genome is analyzed. In a metagenomic sample certain genomes have a small number of reads mapped to them making this approach inapplicable for GC normalization of actual data. On the other hand, it is possible to create an approximation model based on a collection of GC bias curves of genomes varying in their GC content (Figure 2). We do not rely on LOESS regression, because it requires large and densely sampled datasets covering the entire two-dimensional parameter space (read GC and genome GC). In our case due to technological limitations we did not have data for very low GC regions of the high-GC genomes or very high GC regions of the low-GC genomes (Figure 2). Therefore, we used multiple nonlinear regression that specified the expected bell-shape curve of the regions with the missing data. Moreover, the resulting regression function can be easily

implemented in a script and applied to metagenomic data generated using the same DNA sequencing setup. It is important to note that the overall GC bias is unique to a particular sequencing setup and comes primarily from the PCR GC bias resulting from the kits used and the downstream instrument GC bias, which is different for different platforms. Therefore, metagenomic labs interested in implementation of our GC bias normalization procedure should collect the data for individual bacterial genomes of varying GC content sequenced in a pooled sample and repeat our regression procedure to identify the regression coefficients specific to their sequencing setup.

Finally, GC normalized reads can be reported per Mb of bacterial reference to account for the increased likelihood of mapping reads to longer genomes. Batch effects can also lead to bias associated with the length of the reference sequence, e.g., in RNA-seq samples. However, it has been reported that that this type of length effect is the strongest in features less than 1000 bp and it plateaus after 5000 bp (14), therefore, it does not affect estimation of bacterial abundances in metagenomic data.

In some cases, metagenomic samples of human flora taken over a certain period of time from the same source, e.g., sputum of a cystic fibrosis patient, afford an opportunity to report changes in bacterial load. In this case, the proportion of human DNA background can be utilized to calculate the total bacterial load, which can reveal the disease stage of the patient. The relative bacterial abundances estimated as described above can be transformed to absolute estimates based on the identified bacterial load.

FUNDING

This work was supported by the Bundesministerium für Bildung und Forschung BMBF (German Center of Lung Research at BREATH, Disease Area Cystic Fibrosis) and the Deutsche Forschungsgemeinschaft (SFB 900, project Z1).

REFERENCES

1. Chen,K. and Pachter,L. (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol*, **1**, 106–112.

2. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

3. Greenblum,S., Turnbaugh,P.J. and Borenstein,E. (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A*, **109**, 594–599.

4. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621–628.

5. Rieber,N., Zapatka,M., Lasitschka,B., Jones,D., Northcott,P., Hutter,B., Jäger,N., Kool,M., Taylor,M., Lichter,P., *et al.* (2013) Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One*, **8**, e66621.

6. Davenport,C.F., Neugebauer,J., Beckmann,N., Friedrich,B., Kameri,B., Kokott,S., Paetow,M., Siekmann,B., Wieding-Drewes,M., Wienhöfer,M., *et al.* (2012) Genometa–a fast and accurate classifier for short metagenomic shotgun reads. *PLoS One*, **7**, e41224.

7. Yakovchuk,P., Protozanova,E. and Frank-Kamenetskii,M.D. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*, **34**, 564–574.

8. Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, **40**, e72.

9. Aird,D., Ross,M.G., Chen,W.-S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, **12**, R18.

10. Romiguier,J., Ranwez,V., Douzery,E.J.P. and Galtier,N. (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*, **20**, 1001–1009.

11. Chandrananda,D., Thorne,N.P., Ganesamoorthy,D., Bruno,D.L., Benjamini,Y., Speed,T.P., Slater,H.R. and Bahlo,M. (2014) Investigating and correcting plasma DNA sequencing coverage bias to enhance aneuploidy discovery. *PLoS One*, **9**, e86993.

12. Miller,C.A., Hampton,O., Coarfa,C. and Milosavljevic,A. (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**, e16327.

13. Risso,D., Schwartz,K., Sherlock,G. and Dudoit,S. (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, 480.

14. Hansen,K.D., Irizarry,R.A. and Wu,Z. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216.

TABLE AND FIGURE LEGENDS

Table 1. Top 30 bacteria found in found in sputum of a cystic fibrosis patient sorted by GC-normalized reads per Mb of reference.

Figure 1. Percentage of filtered out reads vs. the total number of mapped reads. Read alignments of thirty cystic fibrosis samples were filtered to remove hits with reads mapped to horizontally transferred genomic islands. There is no significant correlation between the sample size and the percentage of filtered out reads.

Figure 2. Normalized coverage vs. GC content of seven bacteria sequenced in a pool. The sequences bacteria are: *Staphylococcus aureus* (dark blue line, 32.8% GC), *Streptococcus pneumoniae* (light blue line, 39.7% GC), *Escherichia coli* (teal line, 50.5% GC), *Klebsiella pneumoniae* (green line, 57.3% GC), *Pandoraea apista* (tan line, 64.9% GC), *Burkholderia cepacia* (orange line, 66.7% GC), and *Nocardia farcinica* (red line, 70.8% GC). Equal amounts

of DNA from each bacterium were sequenced in a pool and mapped to their respective genome references. The normalized coverage was calculated for each GC percentage point based on the proportion of the number of reads mapped to 100 bp genome windows having this GC content to the number of such windows and normalized by the proportion of the expected number of all reads to the number of all windows for a given genome.

Figure 3. Contour plot of the proposed GC normalization model. The model approximates expected normalized coverage of genomic regions stratified by their GC content for bacterial genomes of various overall GC content in a pooled whole-metagenome shotgun sequencing sample. Generally, the regions higher than the contour line of value 1 are overrepresented and the regions lower than this line are underrepresented. To perform the GC normalization, each binned read in a metagenomic sample should be divided by the normalized coverage value predicted based on its GC content and the content of the genome to which it mapped.

Figure 4. Relative and absolute abundances of bacteria in upper airways of two cystic fibrosis patients. The two upper graphs show relative (left) and absolute (right) bacterial abundance estimates of a patient with the homozygous F508del mutation in the CFTR gene. The three samples were taken roughly three months apart and reflect disease progression and response to treatment. Changes in the amount of *Pseudomonas aeruginosa* characteristic of the cystic fibrosis disease grade can be clearly observed in the absolute abundances graph, while the relative abundances can be misleading. The two lower graphs show the data for a patient with the 1898+3 A-G mutation in the CFTR gene. This is a much milder case compared with the previous patient. This is only evident from the absolute abundances graph showing much lower bacterial loads in the upper airways of the patient.

Figure 1

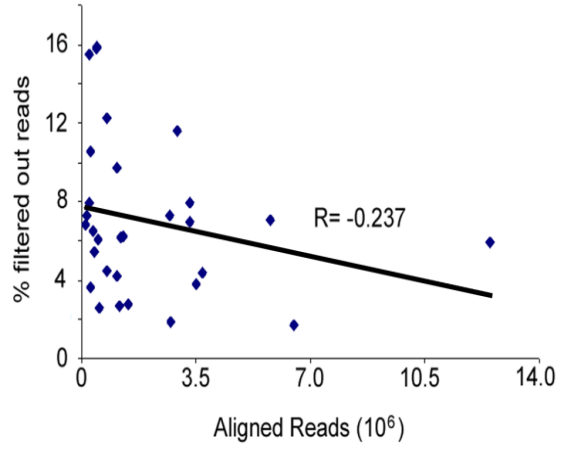


Figure3

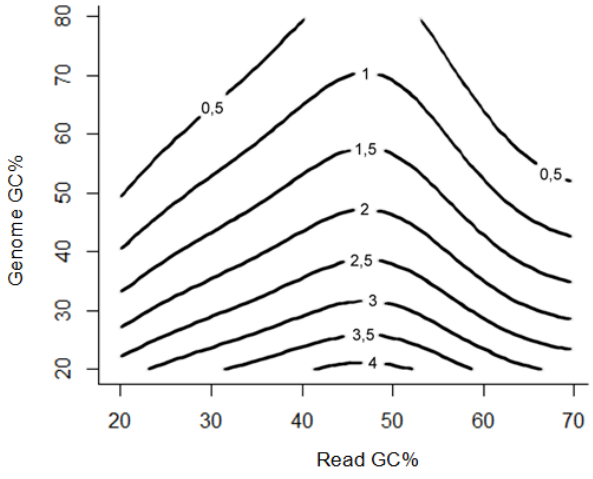


Figure 2

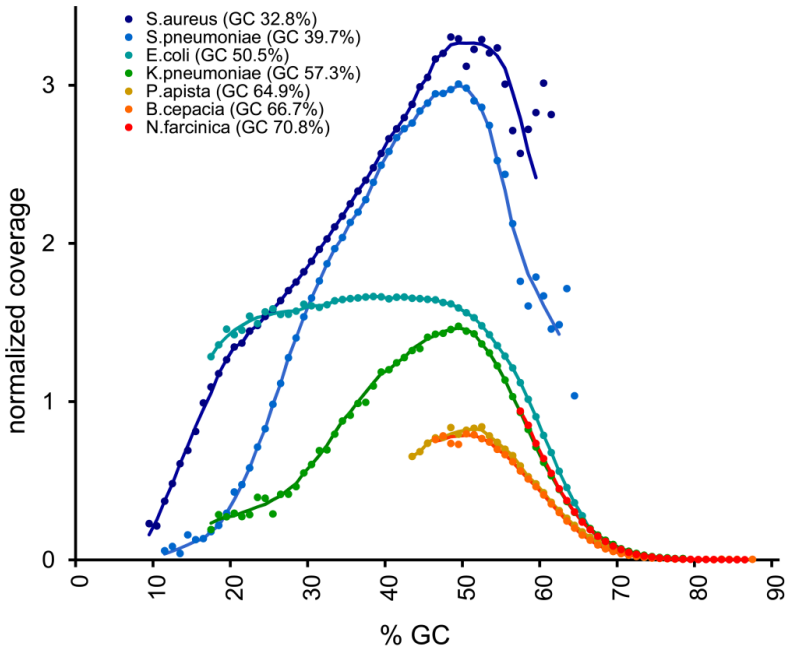
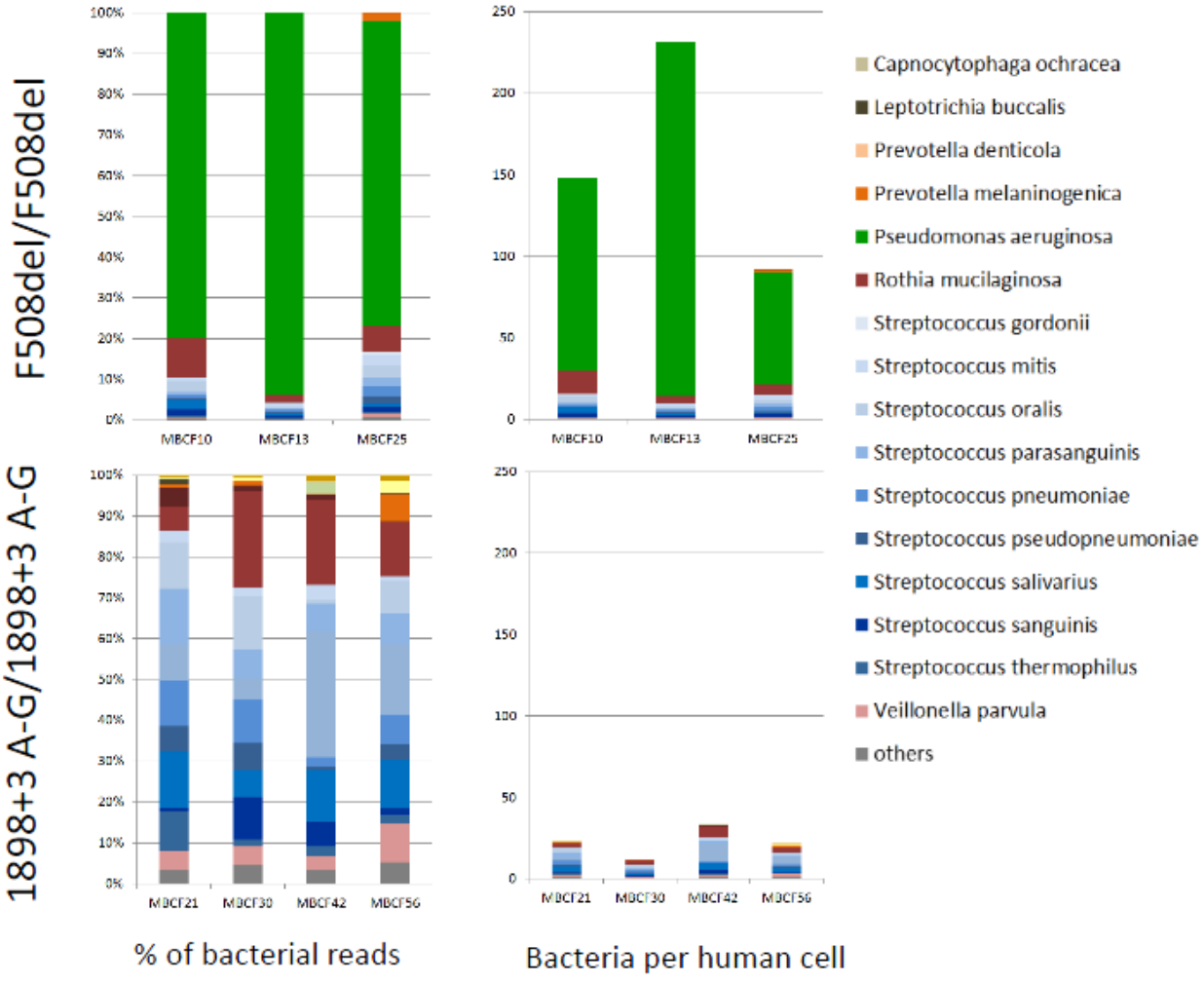


Figure 4



Chapter 7

The cystic fibrosis lower airways microbial metagenome

7.1 Background

Whole genome shotgun (WGS) sequencing has been identified as the most robust and comprehensive method for metagenomics research, becoming an alternative to 16S rDNA sequencing. Although, the majority of microbial community studies have utilized 16S rDNA sequencing techniques, WGS provides a deeper analysis making the identification of low taxonomic levels, like species or strains, to one of its major advantages compared to previous technologies. WGS sequencing provides further benefits such as the quantification of absolute abundances of species, reconstruction of metabolic pathways, gene prediction or identification and reconstruction of novel genomes[90].

WGS sequencing projects produce vast amounts of data which makes the computational and bioinformatics analysis to one of the main keys for a successful metagenomic study. Numerous tools have been developed to handle 454 pyrosequencing or Illumina metagenome sequences, but there are only few computational resources for the analysis of metagenome sequences generated by the SOLiD platform.

WGS metagenome sequencing has already been applied to explore microbial communities in human habitats, for example, in skin[91], intestine[6,29,92] and in the global human microbiome project[10].However, a comprehensive and exhaustive analysis of the cystic fibrosis lower airways microbiome has not been done before.

Cystic fibrosis (CF) is the most common lethal autosomal recessive disorder in Caucasians populations. It is caused by mutations in the *CFTR* gene which encodes the CF transmembrane conductance regulator (CFTR). The major clinical manifestations are severe pulmonary and intestinal symptoms, in particular chronic pulmonary inflammation, microbial lung infections, intestinal obstruction and pancreatic insufficiency. CFTR transports chloride and bicarbonate

transport across the apical membrane of epithelial cells. CFTR is defective or absent in CF. Consequently an acidified and dehydrated milieu is generated in the extracellular epithelial lumen which provokes mucus plugging and obstruction of ducts. In the airways this thick mucus provides a favorable environment for the colonization and growth of opportunistic bacterial pathogens causing inflammatory responses and infections in CF patients. For this reason, a better understanding of the complexity of the cystic fibrosis lower airways microbial community is needed as well as the identification of the nucleotide variants in the main pathogens or antibiotic resistance genes.

7.2 About the paper

In this article, we present the first comprehensive and unbiased study of the cystic fibrosis lower airways metagenome. We determined the composition of viruses, fungi and bacteria of temporal series of induced sputa collected from 15 exocrine pancreas insufficient (PI) and 10 exocrine pancreas sufficient (PS) individuals with CF. To perform the analyses we used direct shotgun 5500xl SOLiD sequencing and a novel in-house pipeline for the analysis of sequences. Our approach incorporated 1800 bacteria, 5804 virus and 610 complete reference genomes downloaded from the NCBI database. Using the new normalization model (described in Chapter 4) for SOLiD technology we identified the absolute and relative abundances of species present in the poly-microbial communities of the CF samples.

Our study demonstrates that a large number of microbial taxa inhabits the CF lower airways. Microbial communities were characterized by an individual signature of multiple lowly abundant species and few CF typical pathogens (like *S. aureus* and *P. aeruginosa*) as the dominant species.

Our approach identified on the average several hundred bacterial taxa and less than 10 DNA viruses or fungi, in all age and disease subgroups.

Author's contribution

In the following study, my contribution was conducting the metagenomic analysis, and also, creating a custom pipeline in perl and R to make this analysis optimized and reproducible.

For Tables, please refer to the DVD attached to the thesis.

The cystic fibrosis lower airways microbial metagenome

Patricia Moran Losada,[1] Philippe Chouvarine,[1] Marie Dorda,[1] Silke Hedtfeld,[1] Samira Mielke,[1] Angela Schulz,[1,2] Lutz Wiehlmann,[1,2] Burkhard Tümmler[1,2*]

[1]Clinic for Paediatric Pneumology, Allergology and Neonatology, OE 6710, Hannover Medical School, Hannover, Germany

[2] Biomedical Research in Endstage and Obstructive Lung Disease (BREATH), Member of the German Center for Lung Research, Hannover, Germany

Abstract

Chronic airway infections determine most morbidity in people with cystic fibrosis (CF). Here we present unbiased quantitative data about the frequency and abundance of DNA viruses, archaea, bacteria, molds and fungi in CF lower airways. Induced sputa were collected on several occasions from children, adolescents and adults with CF. Deep sputum metagenome sequencing identified on average about ten DNA viruses or fungi and several hundred bacterial taxa. The metagenome of a CF patient was typically found to be made up of an individual signature of multiple lowly abundant species superimposed by few disease-associated pathogens such as *Pseudomonas aeruginosa* and *Staphylococcus aureus* as major components. The host-associated signatures ranged from inconspicuous poly-microbial communities in healthy subjects to low-complexity microbiomes dominated by the typical CF pathogens in patients with advanced lung disease. The DNA virus community in CF lungs mainly consisted of phages and occasionally of human pathogens such as adeno- and herpesviruses. The *S. aureus* and *P. aeruginosa* populations were composed of one major and numerous minor clone types. The rare clones constitute a low copy genetic resource which could rapidly expand as a response to habitat alterations such as antimicrobial chemotherapy or invasion of novel microbes.

Introduction

Cystic fibrosis (CF) is a life-shortening, debilitating, autosomal recessive disease that is caused by mutations in the *Cystic Fibrosis Transmembrane Conductance Regulator* (*CFTR*) gene [1]. The basic defect of impaired epithelial chloride and bicarbonate secretion predisposes to chronic airway infections with opportunistic pathogens which determine most morbidity in people with CF [1]. Epidemiological data drawn from culture-dependent diagnostics of respiratory specimens indicated that CF patients become colonized in their airways with *Haemophilus influenzae* and *Staphylococcus aureus* during early childhood followed by *Pseudomonas aeruginosa* and sometimes by organisms such as *Burkholderia cepacia* complex or atypical mycobacteria later in life [2]. Culture-independent technologies, however, revealed that the CF respiratory tract is not inhabited by these few pathogens, but rather by complex poly-microbial communities [3 – 10]. Sequencing of PCR-amplified parts of bacterial 16S rDNA genes could identify over 100 distinct genera including *Streptococcus* and numerous anaerobes as major players that are routinely not detected during the culture-dependent processing of CF-derived respiratory secretions.

Sequence variations in the ancient and ubiquitous ribosomal RNA genes are considered to reflect the universal molecular clock of life [11]. Correspondingly the composition of microbial communities is described by the abundance of individual rDNA sequences. Yet one has to accept some inherent biases of this approach. First, prokaryotes and eukaryotes need to be analyzed separately due to their basic difference of rDNA sequence and correspondingly most work has been confined to the bacterial microbiota. Second, the sequences of the sample are prepared for analysis by oligonucleotide primer-based amplification steps. Subsequent sequencing of the amplicons has a limited ability to resolve taxonomic identification to the species level, may fail to detect phyla and can skew the estimation of species relative abundance in a community [12 - 13].

These limitations can be overcome by whole-genome shotgun sequencing (WGS) [14] that allows a functional assessment of the gene content of the community and may provide information about the composition of the community up to the level of clonal complexes. Within the context of the CF lungs, the published studies have so far focused on the sputum metagenomes derived from CF adults with advanced lung disease [15 – 17]. Each of the ten so far investigated subjects hosted a unique poly-microbial community.

This study extends the scope to patients of all age groups and grades of disease severity in order to cover the whole range of lower airways microbial metagenomics in CF. Induced sputa collected from exocrine pancreatic sufficient (PS) and exocrine pancreatic insufficient (PI) children, adolescents and adults with CF were investigated by WGS in order to identify the bacteria, archaea, DNA viruses, molds and fungi residing in the respiratory secretions. Normalization provided quantitative data about the relative and absolute abundance of microbial

species. Moreover, by focusing on the major CF pathogens *S. aureus* and *P. aeruginosa* the metagenome sequence data sets were examined for the number of co-existing clone types and uncommon, probably de novo mutations in antimicrobial resistance determinants. In the future such a metagenome-guided deep insight into clone and sequence variations could assist in our management of respiratory tract infections.

Methods

Patients. Subjects with CF were recruited from the CF clinic of Hanover Medical School. All patients had been regularly seen at the CF clinic since the age of diagnosis. The diagnosis of CF had been made by the detection of two disease-causing mutations in the *CFTR* gene [18] and elevated chloride concentrations in the Gibson-Cooke pilocarpine iontophoresis sweat test [19] or a Sermet score in the CF range of nasal transepithelial potential difference measurements [20, 21] and/or chloride secretory responses in the CF range of intestinal current measurements [22]. Exocrine pancreatic status was assessed by the fecal-elastase-1 test [23]. Lung function was assessed by spirometry, bodyplethysmography and in the healthier subjects by multiple breath nitrogen washout [24, 25]. The 15 exocrine pancreas insufficient (PI) patients were homozygous for the most common CF mutation p.Phe508del [18]. The 10 exocrine pancreas sufficient (PS) subjects were either compound heterozygous for a PI- and a PS-conferring *CFTR* mutation [18] (9 subjects) or homozygous for a PS mutation [18]. At the date of recruitment patients were either 8 – 13 years old (children, group A), 18 – 23 years old (adolescents and young adults, group B) or elder than 28 years (adults, group C). Patients were also classified by disease severity. Healthy CF subjects had normal anthropometry (body mass index > 19) and normal lung function, i.e. multiple breath nitrogen washout revealed a normal lung clearance index and spirometry yielded FEV1 values of more than 90% predicted. Mildly affected CF patients (category ‘mild’) exhibited normal anthropometry, an anomalous lung clearance index and FEV1 values of 70% - 110% predicted at the day of recruitment. Lung function was chronically compromised for three years or more in all moderately or severely affected CF patients (FEV1 50 - 70% predicted, category ‘moderate’; FEV1 30 – 50% predicted in the absence of an acute pulmonary exacerbation, category ‘severe’; FEV1 < 30% predicted in the absence of an acute pulmonary exacerbation, category ‘end-stage lung disease’). The study was approved by the Ethics Committee of Hannover Medical School (no. 1510-2012).

Wet-lab experimental procedures. The procedures are described in the supplementary material. Briefly, induced sputum was collected by autogenic drainage during cycles of 3-minutes inhalation of 3% hypertonic saline. After sputa had been diluted with buffer and subjected to hypotonic lysis, DNA was purified from the suspension according to the ‘Hard-to-lyse-Bacteria’ protocol with the NucleoSpin Tissue kit (Machery-Nagel, Düren). DNA libraries were prepared from sheared DNA according to an in-house protocol. Sequencing was performed on a SOLiD 5500XL system (Life Technologies) in color space with 75 bp read length and implemented Exact call chemistry (Life Technologies).

In silico analyses. Approaches and software are described in the supplementary material. In brief, the raw sequence reads were trimmed and then (in this order) low quality reads, human reads, non-human low-complexity reads and non-human reads encoding mobile genetic elements were removed. The remaining microbial reads were normalized by GC content and genome length. This curated data set was then used for the identification of taxa (DNA viruses, bacteria, archaea, molds and fungi), principal component analysis, search for mutation in antimicrobial resistance genes and analysis of the *S. aureus* and *P. aeruginosa* populations in the respiratory secretions.

Results

Lower airways microbial metagenome of PS and PI individuals with CF

Induced sputa were collected on several occasions from 10 PS and 15 p.Phe508del homozygous PI subjects with CF. Deep metagenome sequencing identified on average less than ten DNA viruses or fungi and several hundred bacterial taxa in samples from children (group A), adolescents (group B) and adults (group C) (Fig. 1b, Fig. S1, data sets in Tables S1 and S2). Bacteria typically made up more than 99% of the microbial community (Fig. 1a). The median contribution of DNA viruses and fungi was on average less than 1%, but it varied between 0.002% and 11% in the individual sample. Anaerobes are characteristic inhabitants of the upper and lower airways of healthy non-CF humans [26]. In our cohort of CF patients the proportion of anaerobes in the microbial metagenome decreased with age (Fig. 1c). This decline started earlier and was more pronounced in PI than in PS subjects consistent with the known earlier onset and more rapid progression of lung disease in p.Phe508del homozygotes who present the typical symptoms of CF disease since birth [1]. The dominant taxa in the cumulative metagenome of the whole cohort were the bacterial species that are most frequently reported from culture-dependent diagnostics, i.e. streptococci, staphylococci, pseudomonads, Haemophilus sp., Burkholderia sp. and Stenotrophomonas sp. (Fig. 1d, Table S2). Thus quantitative metagenomics fits with culture-based epidemiological data for the most abundant bacterial species in CF lungs.

Figure 2 provides a detailed overview of the frequency of recovery of all detected species and their relative abundance. Taxa and primary data are listed in Tables S1 and S3. The viral community consisted primarily of phages (Fig. 1d), a few human pathogens, primarily herpes virus and adenovirus, and rare cases of viruses infecting non-mammalian eukaryotic hosts. Dominant species in the mycobiome were Aspergillus species and Saccharomycetes including Candida sp. consistent with current knowledge of CF mycology [27]. The community of bacteria and archaea turned out to be highly diverse including numerous species and phyla that yet have not been reported to inhabit the niche of the human CF lung. The lead pathogen *P. aeruginosa* was identified in all specimens from PI patients although six of them had been classified as *P. aeruginosa* negative according to the clinical records. This finding suggests the ubiquitous

presence of *P. aeruginosa* in respiratory secretions of PI CF patients. However, in all specimens taken from these misclassified patients *P. aeruginosa* was only present at low abundance on average of 0.02% of all reads.

The cladograms in Figure 3 focus on the bacteria which made up the top 95% of the cumulative metagenome population of PI (Fig. 3a) and PS CF patients (Fig.3b). The population in PI CF airways was dominated by pseudomonads and staphylococci followed by Veillonella, Streptococci, Prevotella, Rothia and other enterobacteriaceae as minor contributors. The spectrum of genera was similar in PS CF patients, but streptococci were more prominent and the population was more diverse and less skewed towards *P. aeruginosa*. Principal component analysis revealed a broadly scattering distribution of data sets retrieved from PS patients and strong clustering of data sets for the samples from PI subjects (Fig. 4) indicating that bacterial communities are more host-specific in PS CF and more disease-specific in PI CF.

Individual microbial metagenome signatures and CF disease severity

Whole metagenome analysis resolved the microbial signature of the individual patient. The spectrum ranged from a normal flora via an intermediate stage when the normal community is perturbed by *H. influenza* or *S. aureus* to a final stage of a low-diversity community dominated by *P. aeruginosa* [28] (Table S1, Fig. 5, Fig. S2). This shift from a normal highly diverse metagenome indistinguishable from that of a healthy subject to the CF-typical end-stage of an almost pure culture of *P. aeruginosa* was correlated in our patient cohort with disease severity, but not with age. As shown in Figure 5, the diversity of the bacterial communities of the top 90% constituents decreased with increasing lung disease severity.

The metagenome of a CF patient was typically found to be made up of an individual signature of multiple lowly abundant species superimposed by few disease-associated pathogens such as *P. aeruginosa* and *S. aureus* as major components. This phenotype became more obvious if we normalized the microbial reads to the human DNA in the sample. For example, as is illustrated in Figure 6, a PS CF patient appeared to have an unrelated metagenome to that of a PI CF patient with chronic colonization with *P. aeruginosa* if presented in bar charts as fractions of total microbial reads (Fig. 6a, 6b). However, after human DNA normalization it can be seen that a quantitatively similar pattern of non-pathogenic species (‘normal flora’) in the two subjects is overshadowed by *P. aeruginosa* in the PI CF patient, whereas no typical CF pathogens are detectable in the healthy PS CF patient (Fig. 6c, 6d). This normalization clarifies some discrepancies in complexity between the microbiomes resolved by comprehensive culture-independent techniques and those based on culture-dependent diagnostics the latter driven to detect disease-associated microbes and to ignore the ‘normal flora’.

Clonal diversity of the *S. aureus* and *P. aeruginosa* populations in CF lungs

The clonal composition of *S. aureus* and *P. aeruginosa* communities in CF sputa was determined from the frequency distribution of SNPs in the metagenomes (Fig. 7, Table S4).

Most *P. aeruginosa* communities consisted of one major clone type (58% to 92% of total; median 83 %), up to two further clones (each contributing to more than 5% of the population) and at least two to nine rare clones (median: five clone types). Similarly, one major clone dominated the *S. aureus* community (53% to 98% of total, median 87%) accompanied by two to seven very minor contributors. In about half of the analyzed samples one or two clones had a share of more than 5%. Previous genotyping and subsequent genome sequencing of serial isolates had suggested that the CF lungs are chronically colonized with co-evolving clades of one or, less frequently, two or three clones [28-31], but our unbiased metagenome data indicate more diverse and more complex compositions of the *S. aureus* and *P. aeruginosa* populations in CF airways.

To identify the genotype of the dominant *S. aureus* and *P. aeruginosa* strains within the frame of published typing schemes, the metagenome sequences were searched for matches with a multi-marker array for *P. aeruginosa* [32] and the MLST database for *S. aureus* [33]. Four of ten analyzed *P. aeruginosa* strains belonged to ubiquitous clones in the global *P. aeruginosa* population [32] and two pairs of 13 *S. aureus* strains were assigned to the common clone type ST7 and the pandemic MRSA lineage ST22 [33], respectively.

Mutations in resistance genes to antimicrobial chemotherapy

The chronic airway infections in CF are treated by chronic or intermittent antimicrobial chemotherapy, at least on the occasion of a pulmonary exacerbation, often accompanied with the emergence of multidrug resistant bacteria as the unwanted side effect [1, 2]. We searched the *S. aureus* and *P. aeruginosa* sequences in the metagenomes for uncommon non-synonymous amino acid substitutions in targets of anti-infectives and/or mediators of antimicrobial resistance (Table 1). Mutations in the *P. aeruginosa* genomes affected genes that are known to be prone for mutation during antipseudomonal chemotherapy [34], but in case of *S. aureus* the mutations also emerged in the gyrase-encoding *gyr* loci [35] that are the targets for fluoroquinolones which the patients' clinicians had never been prescribed as antistaphyloccocal chemotherapy. Besides the improbable cross-infection with a resistant strain the treatment of concomitant infections by *P. aeruginosa* with a fluoroquinolone would be the most likely explanation for the collateral mutations in the *S. aureus gyr* genes. This example demonstrates the power of non-selective metagenome sequencing to detect genetic variations in traits of interest such as drug resistance, virulence or, lifestyle.

Discussion

Deep metagenome sequencing revealed a large repertoire of viruses, molds, fungi, archaea and bacteria in the CF lung habitat. The lower airways metagenome of a CF patient was typically found to be made up of an individual signature of multiple lowly abundant species superimposed by few classical CF pathogens such as *P. aeruginosa* and *S. aureus* as major components. This phenotype became more obvious if we normalized the microbial reads to the

human DNA in the sample. This presentation of data in terms of absolute abundance of microbes as shown in Figure 6 was more similar to the outcome of routine culture-dependent diagnostics that just communicate disease-associated aerobes, than the common output format of microbiomes as shown in Figure 5 that normalizes all patients' samples to 100% irrespective of the absolute contents of microbes in the respiratory secretions [3 – 10]. In other words, the outcome of culture-dependent and culture-independent analysis of CF respiratory specimens is more similar than discordant modes of presentation may suggest.

The paediatric CF microbiome has been shown to be more diverse than that of CF adults indicating that there may be a time window for therapeutic intervention that maintain diversity while reducing total bacterial load [7, 36]. Our study now shows cases of young PI or PS CF adults who still have a healthy microbial metagenome. All study participants have been regularly seen since the age of diagnosis by a dedicated team of CF caregivers suggesting that continuous surveillance and intervention, if indicated, could prevent or decelerate progressive CF lung disease in some, but not all subjects.

Consistent with literature reports [37, 38] the lead CF pathogen *P. aeruginosa* became a major member of the microbial community in subjects with compromised lung function. Mucus plugging, airway remodeling, micro-colony and biofilm formation will then drive the regional isolation of the microbial metagenome [28, 39]. Considering this spatial heterogeneity, all sputa were collected by autogenic drainage in order to retrieve metagenomes that are representative for the whole lung.

Within-clone evolution of major clones is thought to trigger the adaptation of *S.aureus* and *P. aeruginosa* to the environment of the CF lungs [17, 28, 30, 31]. Our metagenome study now demonstrates that this concept does not cover the whole scenario. The *S. aureus* and *P. aeruginosa* populations do not only consist of one to three major clones, but also of numerous rare clones. These infrequent clones constitute a low copy genetic resource which could rapidly expand as a response to habitat alterations such as antimicrobial chemotherapy or invasion of novel microbes. Thanks to the high accuracy of sequencing-by-ligation in the color space of 99.943% we could resolve the clonal diversity of *S. aureus* and *P. aeruginosa* in CF airways. The error rates of the more often used sequencing-by-synthesis or single molecule real-time sequencing technologies are too high to reliably detect infrequent sequence variants which may explain why minor constituents of poly-microbial communities at the rank of strains and clone types have yet not been reported in the literature.

Metagenome sequencing generated quantitative and unbiased data about microbial diversity in CF lungs. Extensive culture-enriched profiling of the CF airway microbiome identified families of bacteria in CF sputa that were not detected by parallel 16S rDNA sequencing [40]. These biases of 16S amplicon sequencing do not apply to metagenome sequencing. Knowing that metagenome sequencing discerned on average one order of magnitude more organisms at the species level than 16S rDNA analyses [40], we consider any culture-enriched molecular

analysis of CF sputum microbes to be dispensable if a metagenome approach is pursued. However, one should bear in mind that the sensitivity of the detection of rare members depends critically on the total number of microbial read sequences (see Fig. S3). Unless one is interested in specific features such as the spectrum of sequence variants in loci of interest, about half a million microbial reads are sufficient to provide a comprehensive metagenome analysis of taxa in CF airways.

References

1. Ratjen F, Bell SC, Rowe SM, Goss CH, Quittner AL, Bush A. Cystic fibrosis. *Nature Dis Primers* 2015; 15010.

2. Waters V, Smyth A. Cystic fibrosis microbiology: Advances in antimicrobial therapy. *J Cyst Fibros* 2015; **14:** 551-560.

3. Surette MG. The cystic fibrosis lung microbiome. *Ann Am Thorac Soc* 2014; **11** Suppl 1: S61-S65.

4. Sibley CD, Parkins MD, Rabin HR, Duan K, Norgaard JC, Surette MG. A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients. *Proc Natl Acad Sci U S A* 2008; **105:** 15070-15075 (2008).

5. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW,Carroll MP, Parkhill J, Bruce KD. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME J* 2011;**5**:780-791.

6. Tunney MM, Klem ER, Fodor AA, Gilpin DF, Moriarty TF, McGrath SJ, Muhlebach MS, Boucher RC, Cardwell C, Doering G, Elborn JS, Wolfgang MC. Use of culture and molecular analysis to determine the effect of antibiotic treatment on microbial community diversity and abundance during exacerbation in patients with cystic fibrosis. *Thorax* 2011;**66**: 579-584.

7. Madan JC, Koestler DC, Stanton BA, Davidson L, Moulton LA, Housman ML, Moore JH, Guill MF, Morrison HG, Sogin ML, Hampton TH, Karagas MR, Palumbo PE, Foster JA, Hibberd PL, O'Toole GA. Serial analysis of the gut and respiratory microbiome in cystic fibrosis in infancy: interaction between intestinal and respiratory tracts and impact of nutritional exposures. *MBio* 2012;**3**(4).

8. Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, Cavalcoli JD, VanDevanter DR, Murray S, Li JZ, Young VB, LiPuma JJ. Decade-long bacterial community dynamics in cystic fibrosis airways. *Proc Natl Acad Sci U S A* 2012; **109**: 5809-5814.

9. Price KE, Hampton TH, Gifford AH, Dolben EL, Hogan DA, Morrison HG, Sogin ML,O'Toole GA. Unique microbial communities persist in individual cystic fibrosis patients throughout a clinical exacerbation. *Microbiome* 2013; **1**: 27.

10. Rogers GB, Shaw D, Marsh RL, Carroll MP, Serisier DJ, Bruce KD. Respiratory microbiota: addressing clinical questions, informing clinical practice. *Thorax* 2015; **70**: 74-81.

11. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB,Euzéby J, Amann R, Rosselló-Móra R. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014; **12**: 635-645.

12. Cai L, Ye L, Tong AH, Lok S, Zhang T. Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. *PLoS One* 2013; **8**: e53649.

13. Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* 2014; **2**: 11.

14. Davenport CF, Tümmler B. Advances in computational analysis of metagenome sequences. *Environ Microbiol* 2013; **15**: 1-5.

15. Lim, Y.W., *et al.* Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J Cyst Fibros* **12**, 154-164 (2013).

16. Lim, Y.W., *et al.* Clinical insights from metagenomics analysis of sputum samples from patients with cystic fibrosis. *J Clin Microbiol* **52**, 425-437 (2014).

17. Quinn, R.A., *et al.* Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *MBio* **5**, e00956-13 (2014).

18. Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, Ramalho AS,Amaral MD, Dorfman R, Zielenski J, Masica DL, Karchin R, Millen L, Thomas PJ,Patrinos GP, Corey M, Lewis MH, Rommens JM, Castellani C, Penland CM, Cutting GR. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet* 2013; **45**: 1160-1167.

19. Gibson LE, Cooke RE. A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine by iontophoresis. *Pediatrics* 1959; **23:** 545-549.

20. Rowe SM, Clancy JP, Wilschanski M. Nasal potential difference measurements to assess CFTR ion channel activity. *Methods Mol Biol* 2011;**741**: 69-86.

21. Sermet-Gaudelus I, Girodon E, Sands D, Stremmler N, Vavrova V, Deneuville E,Reix P, Bui S, Huet F, Lebourgeois M, Munck A, Iron A, Skalicka V, Bienvenu T,Roussel D, Lenoir G, Bellon G, Sarles J, Macek M, Roussey M, Fajac I, Edelman A. Clinical phenotype and genotype of children with borderline sweat test and abnormal nasal epithelial chloride transport. *Am J Respir Crit Care Med* 2010; **182**: 929-936.

22. Derichs N, Sanz J, Von Kanel T, Stolpe C, Zapf A, Tümmler B, Gallati S,Ballmann M. Intestinal current measurement for diagnostic classification of patients with questionable cystic fibrosis: validation and reference data. *Thorax* 2010; **65:** 594-599.

23. Leeds JS, Oppong K, Sanders D.S. The role of fecal elastase-1 in detecting exocrine pancreatic disease. *Nat Rev Gastroenterol Hepatol* 2011; **8:** 405-415.

24. Jensen R, Stanojevic S, Gibney K, Salazar JG, Gustafsson P, Subbarao P, Ratjen F. Multiple breath nitrogen washout: a feasible alternative to mass spectrometry. *PLoS One* 2013; **8**: e56868.

25. Jensen R. Standard operating procedure: Multiple breath nitrogen washout. Version 1.0. Toronto, 2013.

26. Develioglu ON, Ipek HD, Bahar H, Can G, Kulekci M, Aygun G. Bacteriological evaluation of tonsillar microbial flora according to age and tonsillar size in recurrent tonsillitis. *Eur Arch Otorhinolaryngol* 2014; **271:** 1661-1665.

27. Chotirmall SH, McElvaney NG. Fungi in the cystic fibrosis lung: bystanders or pathogens? *Int J Biochem Cell Biol* 2014; **52:**161-173.

28. Jorth P, Staudinger BJ, Wu X, Hisert KB, Hayden H, Garudathri J, Harding CL,Radey MC, Rezayat A, Bautista G, Berrington WR, Goddard AF, Zheng C, Angermeyer A, Brittnacher MJ, Kitzman J, Shendure J, Fligner CL, Mittler J, Aitken ML,Manoil C, Bruce JE, Yahr TL, Singh PK. Regional isolation drives bacterial diversification within cystic fibrosis lungs. *Cell Host Microbe* 2015; **18**: 307-319.

29. Cramer N, Wiehlmann L, Ciofu O, Tamm S, Høiby N, Tümmler B. Molecular epidemiology of chronic *Pseudomonas aeruginosa* airway infections in cystic fibrosis. *PLoS One* 2012; **7**: e50731.

30. Cramer N, Klockgether J, Wrasman K, Schmidt M, Davenport CF, Tümmler B. Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. *Environ Microbiol* 2011; **13**: 1690-1704.

31. Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genet* 2013; **9:** e1003741.

32. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Köhler T, van Delden C, Weinel C, Slickers P, Tümmler B. Population structure of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 2007; **104**: 8101-8106.

33. 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.

34. Poole, K. Efflux-mediated multiresistance in Gram-negative bacteria. *Clin Microbiol Infect* 2004; **10:** 12-26.

35. Coskun-Ari FF, Bosgelmez-Tinaz G. *grlA* and *gyrA* mutations and antimicrobial susceptibility in clinical isolates of ciprofloxacin- methicillin-resistant *Staphylococcus aureus. Eur J Med Res* 2008; **13:** 366-370.

36. Hampton TH, Green DM, Cutting GR, Morrison HG, Sogin ML, Gifford AH, Stanton BA, O'Toole GA. The microbiome in pediatric cystic fibrosis patients: the role of shared environment suggests a window of intervention. *Microbiome* 2014; **2**:14.

37. Blainey PC, Milla CE, Cornfield DN, Quake SR. **Quantitative analysis of the human airway microbial ecology reveals a pervasive signature for cystic fibrosis.** *Sci Transl Med* 2012; **4:** 153ra130.

38. Stressmann FA, Rogers GB, van der Gast CJ, Marsh P, Vermeer LS, Carroll MP, Hoffman L, Daniels TW, Patel N, Forbes B, Bruce KD. **Long-term cultivation-independent microbial diversity analysis demonstrates that bacterial communities infecting the adult cystic fibrosis lung show stability and resilience.** *Thorax* 2012; **67:** 867-873.

39. Boon M, Verleden SE, Bosch B, Lammertyn EJ, McDonough JE, Mai C, Verschakelen J, Kemner-van de Corput M, Tiddens HA, Proesmans M, Vermeulen FL, Verbeken EK,Cooper J, Van Raemdonck DE, Decramer M, Verleden GM, Hogg JC, Dupont LJ,Vanaudenaerde BM, De Boeck K. Morphometric analysis of explant lungs in cystic fibrosis. *Am J Respir Crit Care Med* 2015 Nov 9.

40. Sibley CD, Grinwis ME, Field TR, Eshaghurshan CS, Faria MM, Dowd SE, Parkins MD, Rabin HR, Surette MG. Culture enriched molecular profiling of the cystic fibrosis airway microbiome. *PLoS One* 2011; **6**: e22702.

Acknowledgments. This study was supported by funds from the Bundesministerium für Forschung und Technologie, German Center for Lung Research (DZL), and the Mukoviszidose e.V. (project 1206). P.M.L. was supported by the Hannover Biomedical Research School (HBRS) and the Center for Infection Biology (ZIB).

Author Contributions. P.M.L. and B.T. conceived the study. M.D., S.H., S.M., A.S. and L.W. performed the experiments. P.M.L. and P.C. wrote scripts. P.M.L., P.C., L.W. and B.T. analyzed the data. The manuscript was written by P.M.L. and B.T.

Author Information. The authors declare no competing financial interests.

Figure Legends.

Figure 1. The cystic fibrosis lower airways metagenome: cumulative presentation of data sets. a, b Box plot presentation of the absolute frequency and the relative abundance (in %) of DNA viruses, bacteria and fungi identified in induced sputa sampled from the lower airways of exocrine pancreas insufficient (PI) and pancreas sufficient (PS) children (A), adolescents (B) and adults (C) with CF. **c,** Box plot presentation of the relative abundance of anaerobic bacteria in the sputa of PI and PS CF children (A), adolescents (B) and adults (C). **d,** The dots in the double logarithmic plot depict the sum of reads identified for bacteriophages and their corresponding bacterial hosts at the genus level. Genera are differentiated by number and color.

Figure 2. Frequency and abundance of microbial species in the cystic fibrosis lower airways metagenome. Each dot in the double logarithmic graphs depicts the identification frequency of a species (in %) and its mean relative abundance (in %) in sputa collected from PI CF (**a, c, e**) and PS CF patients (**b, d, f**). Abundance was separately normalized for bacteria (**a, c**), DNA viruses (**c, d**), molds and fungi (**e, f**). The name of a taxon is depicted for the most common and most abundant species. Color codes: (**a, b**) Bacterial taxa belonging to the same phylum are indexed by matching color. (**c, d**) Viruses are differentiated by their host. (**e, f**) Molds and fungi are differentiated by class.

Figure 3. Taxonomic cladograms reporting the most abundant species contributing to the top 95% of the bacterial communities of the lower airways of PI (**a**) and PS (**b**) patients with CF. Circle size is proportional to the log of average abundance. The height of the segments of the outermost ring indicates the relative abundance of the respective genera and phyla in all age groups. The next three outermost circles indicate the relative abundance of clades by color intensity for adults (green), adolescents (purple) and children (olive). The analysis is based on 38 samples from PI and 24 samples from PS patients. Species of the genus streptococcus are a, *S. parasanguinis*; b, *S. salivarius*; c, *S. oralis*; d, *S. mitis*; e, *S. pneumoniae*; f, *S. thermophilus*; g, *S. sanguinis*; h, *S. pseudopneumoniae*; I, *S. gordonii*.

Figure 4. Principal component analysis of the sputum metagenome data sets of PI (red) and PS (red) patients either normalized (lower panel, **d, e, f**) or not normalized (upper panel, **a, b, c**) to human DNA in the sample.

Figure 5. Composition of microbial communities in induced sputa of (from left to right) healthy, mildly, moderately, severely or very severely (end-stage lung disease) affected individuals with CF. Data are presented in a stacked bar chart of relative abundance as fractions of total reads (*y*-axis) of the top 90% of species for each patient (*x*-axis). The color key indicating these species is shown at the right hand side of the figure.

Figure 6. Bacterial sputum microbiomes of a PI CF patient who is homozygous for the most common *CFTR* mutation p.Phe508del (a, c) and of a PS CF patient (b, d) who is

homozygous for the rare splice mutation in the *CFTR* gene c.1766+3 A-G. The same data sets are either presented as stacked bar charts of relative abundance as fractions of total bacterial reads (**a, b**) or as number of bacterial reads per human cell (**c, d**). Whereas the normalized bacterial microbiomes (**a, b**) suggest unrelated compositions of the bacterial communities in the respiratory tracts of the two patients, the normalization per human cell unravels similar colonization modes with anaerobes in the two patients which are overshadowed by the dominant *P. aeruginosa* community in the PI patient. The tracings at the top of the figures show original recordings of short circuit currents in rectal suction biopsies to characterize the basic defect and CFTR function in the two individuals with CF. The horizontal double-headed arrow corresponds to a recording time of 30 min and each vertical double-headed arrow to a current of 10 µA/cm^2. Intestinal current measurements (ICM) were performed per protocol by the addition of amiloride (1), indomethacin (2), carbachol (3), cAMP/forskolin (4), DIDS (5) and histamine (6). Carbachol, cAMP/forkolin and histamine induce chloride secretory responses seen in the ICM by signals in the upward direction. The tracings of the two patients show either no (**a, c**) or intermediate CFTR activity (**b, d**) as indicated by the cumulative chloride currents of 0 or 24 µA/cm^2 implying that no functional CFTR is operating in the F508del homozygote whereas substantial residual CFTR activity is present in the PS CF subject.

Figure 7. Clonal diversity of the *S. aureus* and *P. aeruginosa* communities in respiratory secretions. Metagenome samples with more than 40,000 species-specific reads were selected for the analysis of SNP diversity in *S. aureus* (blue) and *P. aeruginosa* (green) sequences at reference genome positions covered by more than 10 reads. The bars show the number of reads (*y*-axis) that at the SNP position were divergent ('mismatch') from the nucleotide of the most prevalent clone ('match'). Bars are sorted in 1%-intervals of the mismatch/match ratio of the reads covering the SNP (*x*-axis). The abundance of the most prevalent clone in the community is given in the upper left corner in terms of the number of reads and its percentage of total reads.

Table 1. Non-synonymous sequence variants of antibiotic resistance genes qualified as rare or *de novo* mutations in *S. aureus* and *P. aeruginosa* populations of CF sputum metagenomes.

Organism	**Gene**	**Amino acid substitution**	**No of patients (no of samples)**
S. aureus	*gyrA*	p.Ser90Leu	2 (4, 2)
		p.Leu576Phe	1 (1)
		p.Arg843His	1 (1)
		p.Thr833fs	1 (1)
		p.Asp891_Glu892dup	1 (2)
	rpoB	p.Asp320Asn	1 (1)
	parE	p.Asn139Ser	2 (2, 1)
		p.Gly530Asp	1 (1)
P. aeruginosa	*mexS*	p.Ala235Thr	1 (3)
		p.Arg48His	1 (1)
	mexI	p.Ala782Glu	2 (1, 1)

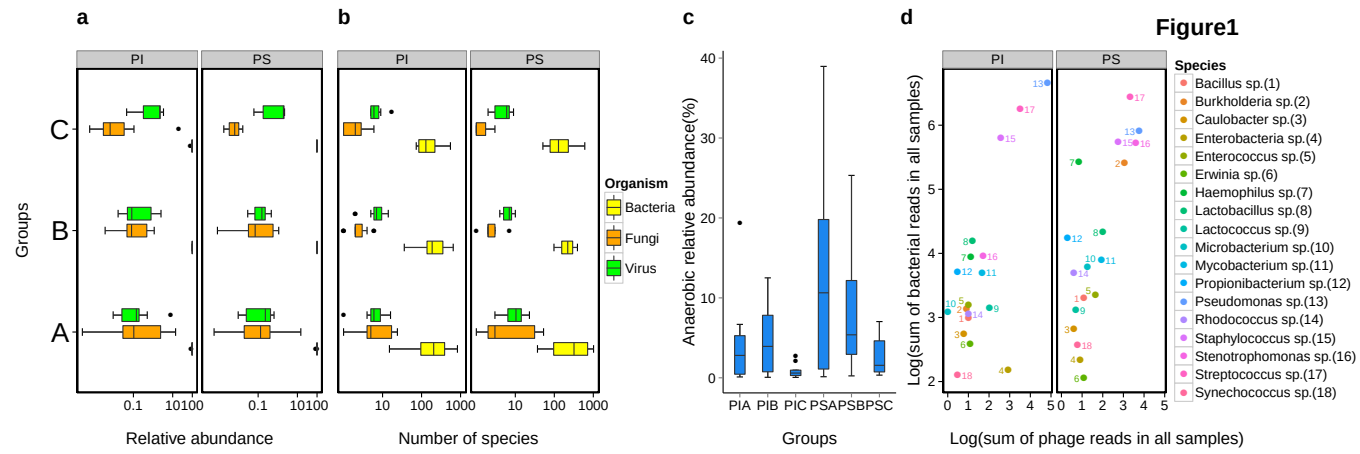
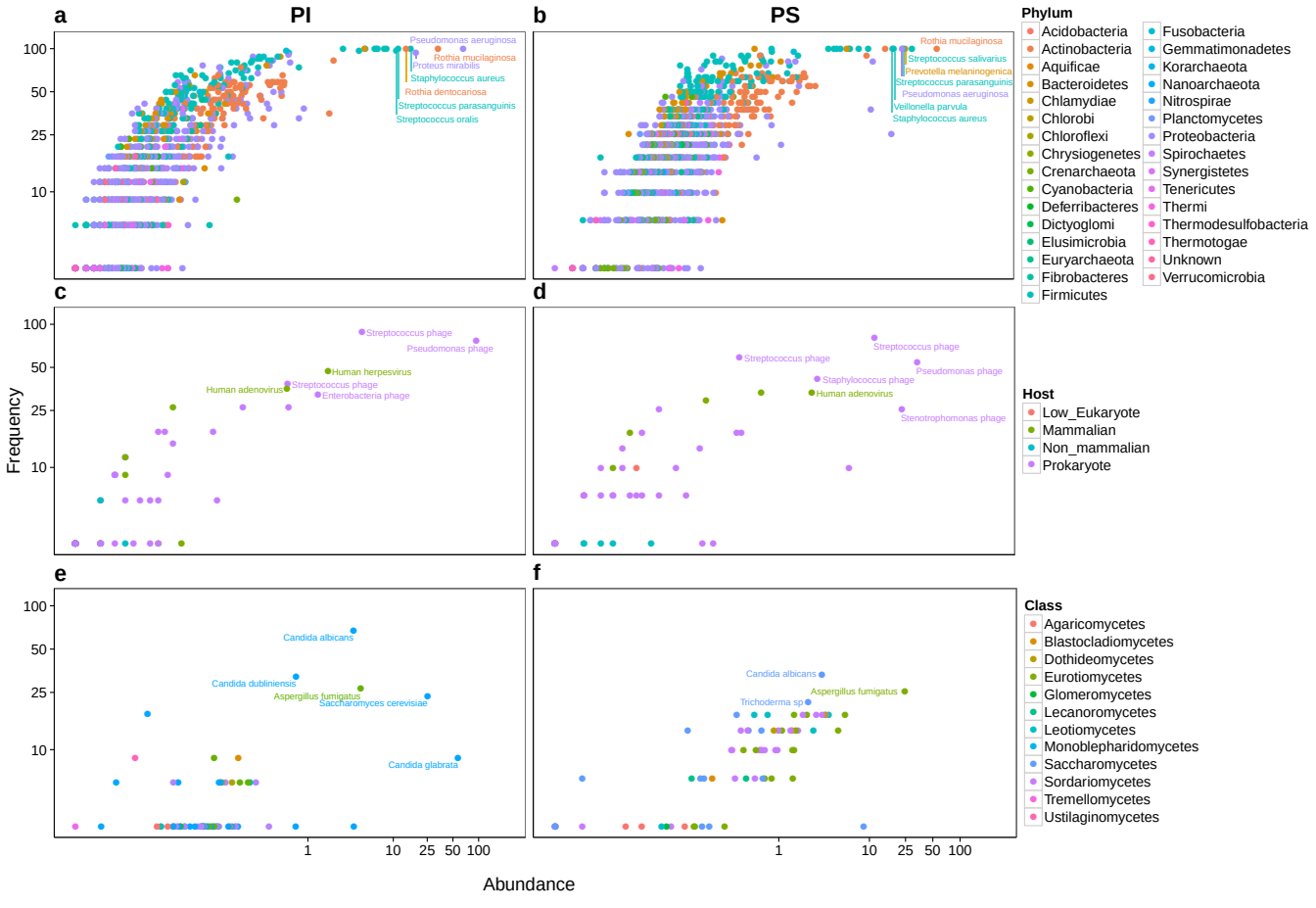


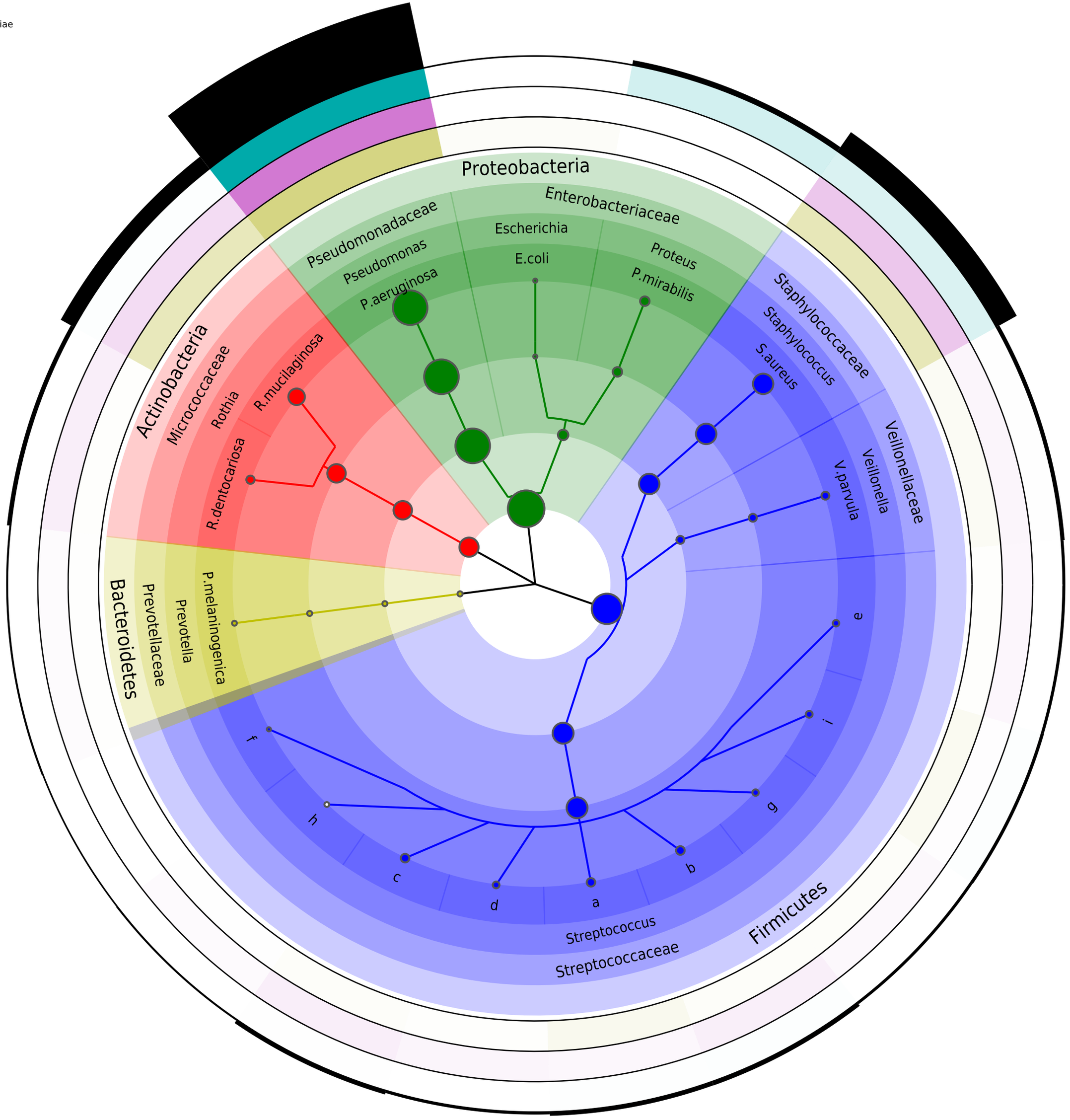
Figure1
a
b
c
d
PI
PS
C
B
A
Groups
Relative abundance
Number of species
Organism
Bacteria
Fungi
Virus
Anaerobic relative abundance(%)
PIA PIB PIC PSA PSB PSC
Groups
Log(sum of bacterial reads in all samples)
Log(sum of phage reads in all samples)
Species
Bacillus sp.(1)
Burkholderia sp.(2)
Caulobacter sp.(3)
Enterobacteria sp.(4)
Enterococcus sp.(5)
Erwinia sp.(6)
Haemophilus sp.(7)
Lactobacillus sp.(8)
Lactococcus sp.(9)
Microbacterium sp.(10)
Mycobacterium sp.(11)
Propionibacterium sp.(12)
Pseudomonas sp.(13)
Rhodococcus sp.(14)
Staphylococcus sp.(15)
Stenotrophomonas sp.(16)
Streptococcus sp.(17)
Synechococcus sp.(18)

Figure 2


a
PI
b
PS
Pseudomonas aeruginosa
Rothia mucilaginosa
Proteus mirabilis
Staphylococcus aureus
Rothia dentocariosa
Streptococcus parasanguinis
Streptococcus oralis
Rothia mucilaginosa
Streptococcus salivarius
Prevotella melaninogenica
Streptococcus parasanguinis
Pseudomonas aeruginosa
Veillonella parvula
Staphylococcus aureus
Phylum
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Chrysiogenetes
Crenarchaeota
Cyanobacteria
Deferribacteres
Dictyoglomi
Elusimicrobia
Euryarchaeota
Fibrobacteres
Firmicutes
Fusobacteria
Gemmatimonadetes
Korarchaeota
Nanoarchaeota
Nitrospirae
Planctomycetes
Proteobacteria
Spirochaetes
Synergistetes
Tenericutes
Thermi
Thermodesulfobacteria
Thermotogae
Unknown
Verrucomicrobia
c
d
Streptococcus phage
Pseudomonas phage
Human herpesvirus
Human adenovirus
Streptococcus phage
Enterobacteria phage
Streptococcus phage
Streptococcus phage
Pseudomonas phage
Staphylococcus phage
Human adenovirus
Stenotrophomonas phage
Host
Low_Eukaryote
Mammalian
Non_mammalian
Prokaryote
Frequency
e
f
Candida albicans
Candida dubliniensis
Aspergillus fumigatus
Saccharomyces cerevisiae
Candida glabrata
Candida albicans
Aspergillus fumigatus
Trichoderma sp
Class
Agaricomycetes
Blastocladiomycetes
Dothideomycetes
Eurotiomycetes
Glomeromycetes
Lecanoromycetes
Leotiomycetes
Monoblepharidomycetes
Saccharomycetes
Sordariomycetes
Tremellomycetes
Ustilaginomycetes
Abundance

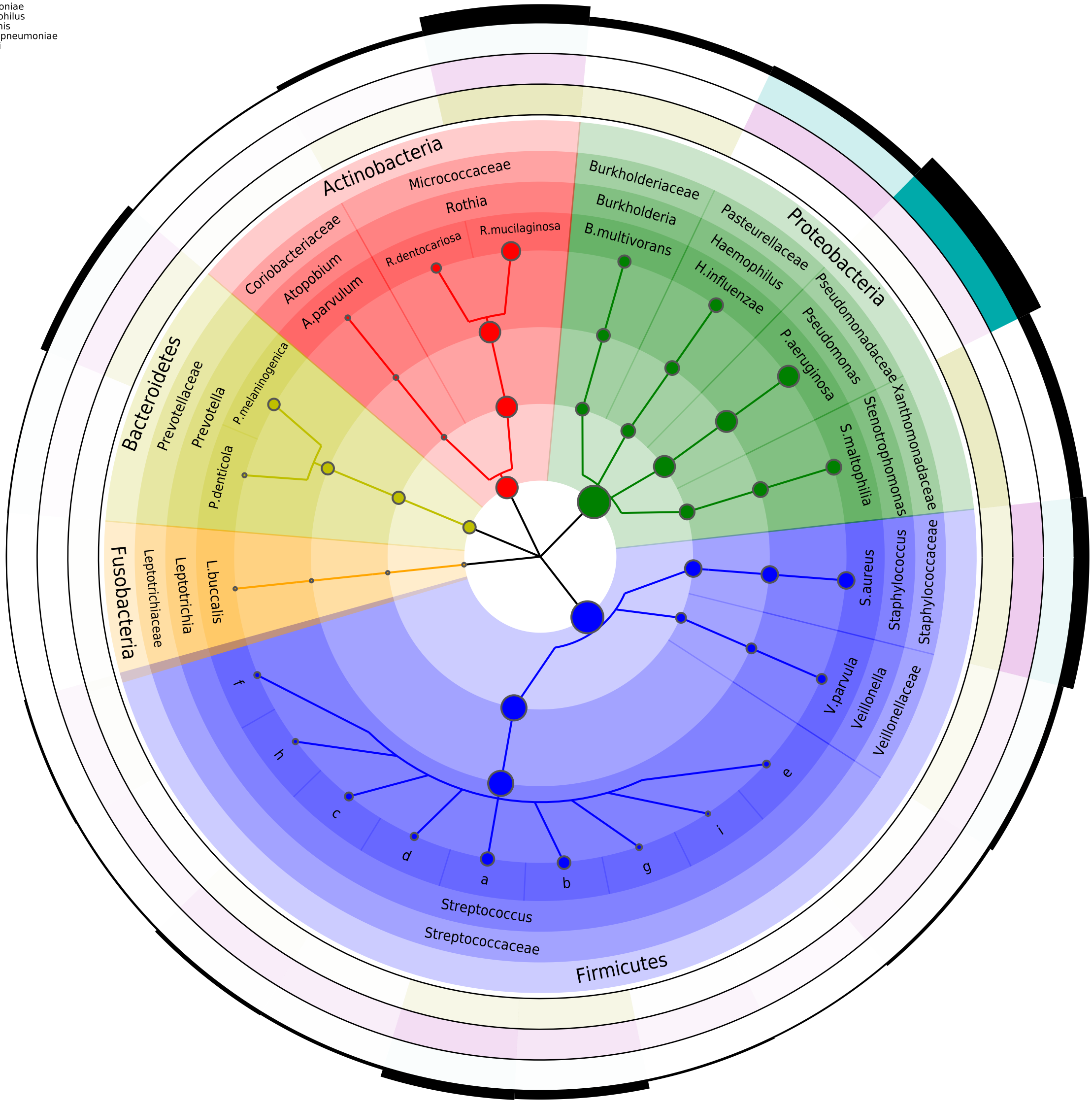
a:S.parasanguinis
b:S.salivarius
c:S.oralis
d:S.mitis
e:S.pneumoniae
f:S.thermophilus
g:S.sanguinis
h:S.pseudopneumoniae
i:S.gordonii

Figure 3a



a:S.parasanguinis
b:S.salivarius
c:S.oralis
d:S.mitis
e:S.pneumoniae
f:S.thermophilus
g:S.sanguinis
h:S.pseudopneumoniae
i:S.gordonii

Figure 3b



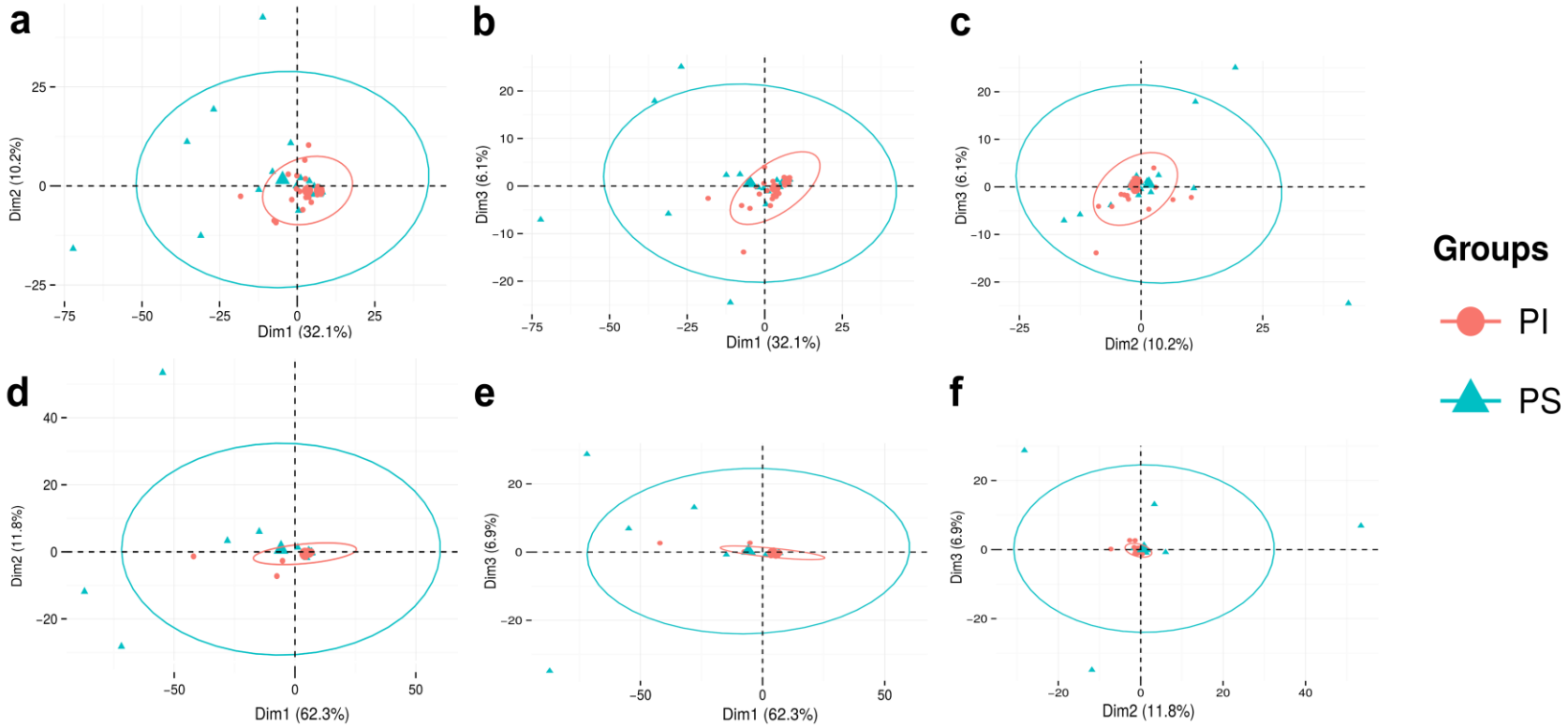
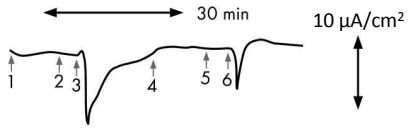

a
b
c
d
e
f
Dim1 (32.1%)
Dim2 (10.2%)
Dim3 (6.1%)
Dim1 (62.3%)
Dim2 (11.8%)
Dim3 (6.9%)
Groups
PI
PS

Figure 4


Healthy
Mild
Moderate
Severe
End−stage
Relative abundances
90
75
50
25
PSBW1.1
PSBW1.2
PSBW1.3
PSBW1.4
PSBM3.1
PSCM2.1
PSBM10.1
PIBM12.1
PIBW11.1
PIBW11.2
PSAW1.1
PSAW1.2
PSAW1.3
PSAW2.1
PSAW2.2
PSAW2.3
PSAW2.4
PIAM2.1
PIAM2.2
PIAM2.3
PIAW11.1
PIBM10.1
PIBM10.2
PIBM10.3
PIBM10.4
PICW11.1
PSAW10.1
PSBW12.1
PIAM5.1
PIAM5.2
PIAM5.3
PIAM5.4
PICM2.1
PICM2.2
PICM2.3
PIAW10.1
PICW7.1
PIAM11.1
PIAM10.1
PIAM12.1
PSAM10.1
PSAM10.2
PSAM10.3
PSCW13.1
PSCW5.1
PSCW5.2
PSCW5.3
PIBM5.1
PIBM5.2
PIBM5.3
PIBM5.4
PICW4.1
PICW4.2
PICW4.3
PICW10.1
PIBM11.1
PIBM11.2
PIBM11.3
PIBM11.4
PIBM11.5
PICM10.1
PICM6.1
Samples
Species
Achromobacter xylosoxidans
Burkholderia ambifaria
Burkholderia cenocepacia
Burkholderia multivorans
Escherichia coli
Haemophilus influenzae
Haemophilus parainfluenzae
Neisseria meningitidis
Proteus mirabilis
Pseudomonas aeruginosa
Pseudomonas mendocina
Stenotrophomonas maltophilia
Atopobium parvulum
Bifidobacterium dentium
Isoptericola variabilis
Kocuria rhizophila
Micrococcus luteus
Rothia dentocariosa
Rothia mucilaginosa
Enterococcus faecalis
Lactobacillus fermentum
Staphylococcus aureus
Streptococcus gordonii
Streptococcus mitis
Streptococcus mutans
Streptococcus oralis
Streptococcus parasanguinis
Streptococcus pneumoniae
Streptococcus pseudopneumoniae
Streptococcus salivarius
Streptococcus sanguinis
Streptococcus thermophilus
Veillonella parvula
Prevotella denticola
Prevotella melaninogenica
Leptotrichia buccalis

Figure 5

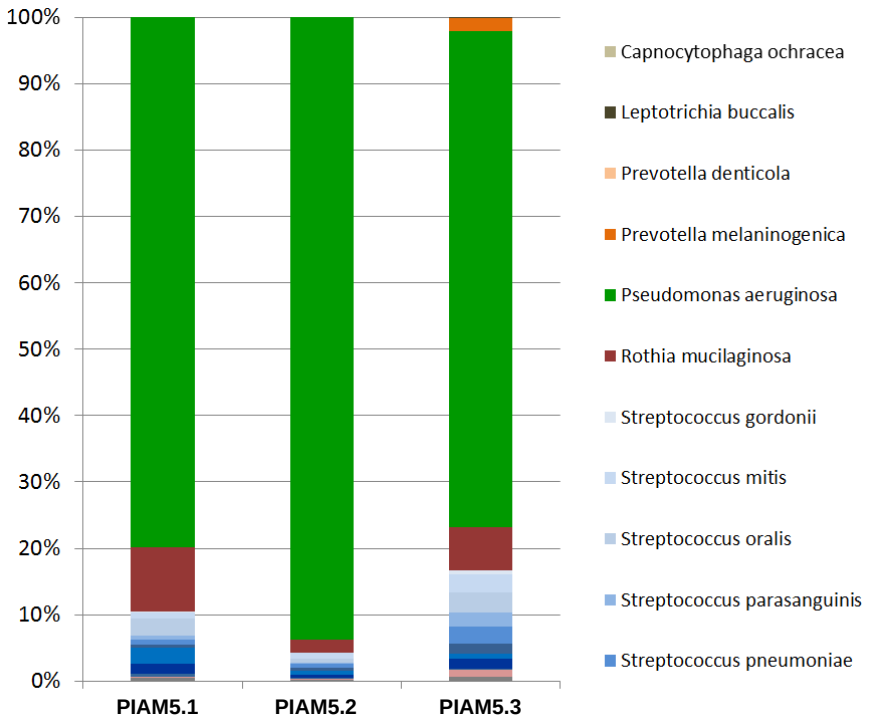
a p.Phe508del/p.Phe508del



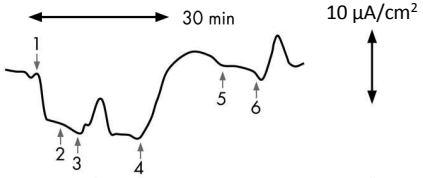
ICM: 0 µA/cm²

Normalization per total bacterial reads

Microbiome:



b c.1766+3 A-G/c.1766+3 A-G



ICM: 24 µA/cm²

Normalization per total bacterial reads

Microbiome:

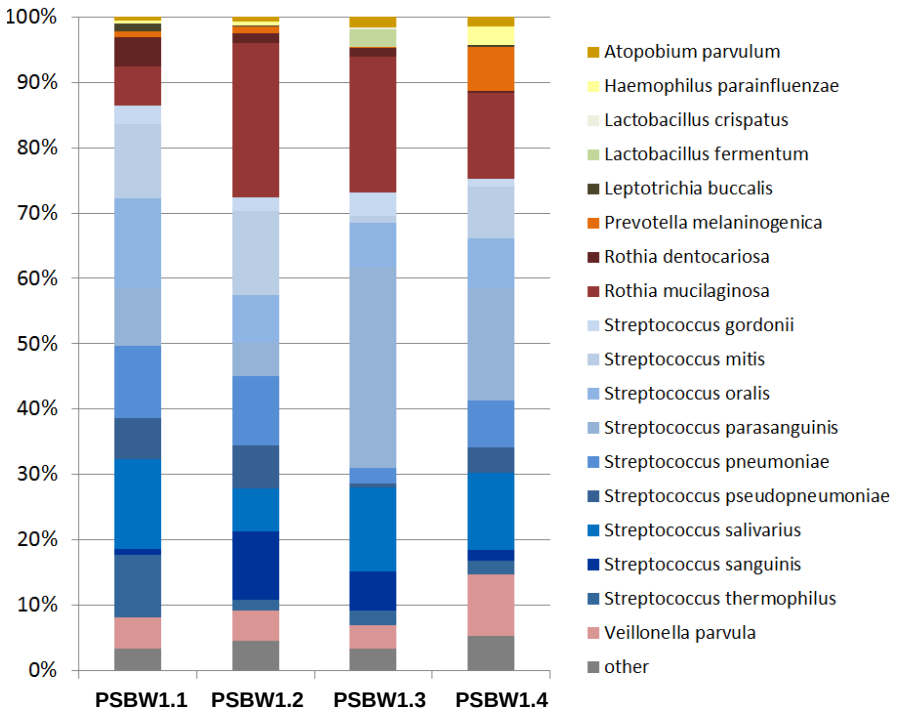
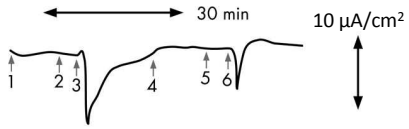


Figure 6

c p.Phe508del/p.Phe508del

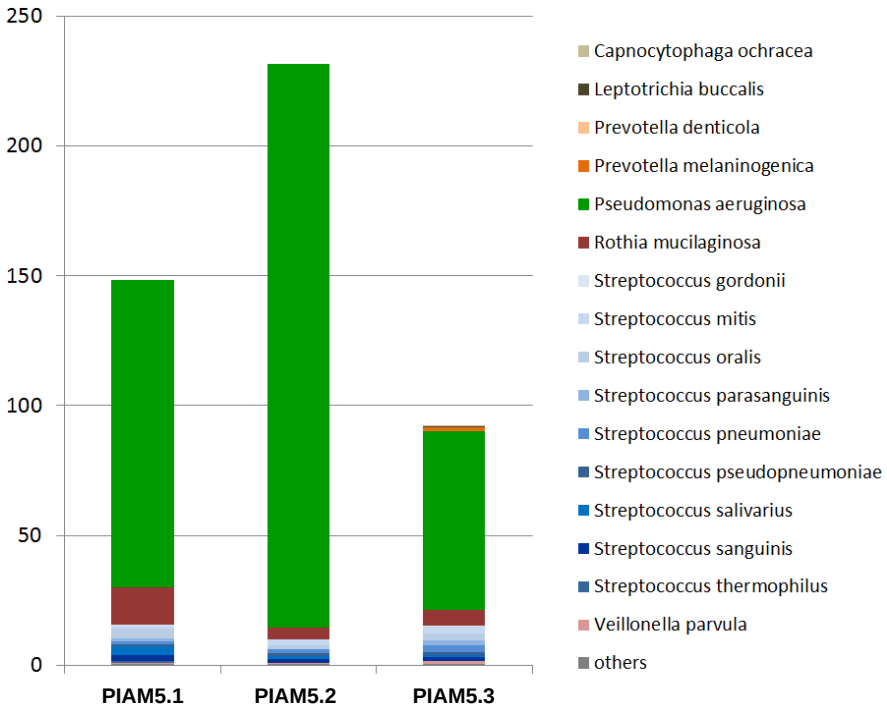


ICM: 0 µA/cm²

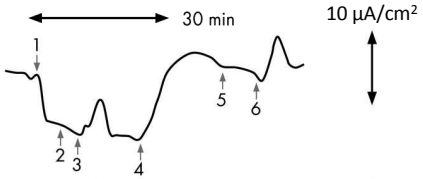
Normalization per human cell

quantitative Microbiome:

bacteria/human cell in sample



d c.1766+3 A-G/c.1766+3 A-G



ICM: 24 µA/cm²

Normalization per human cell

quantitative Microbiome:

bacteria/human cell in sample

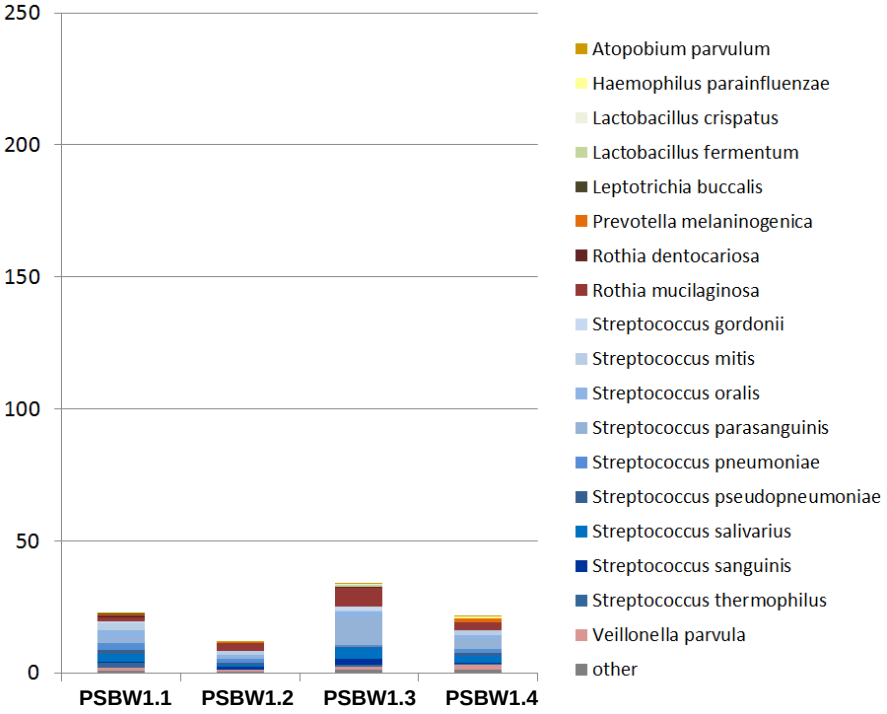
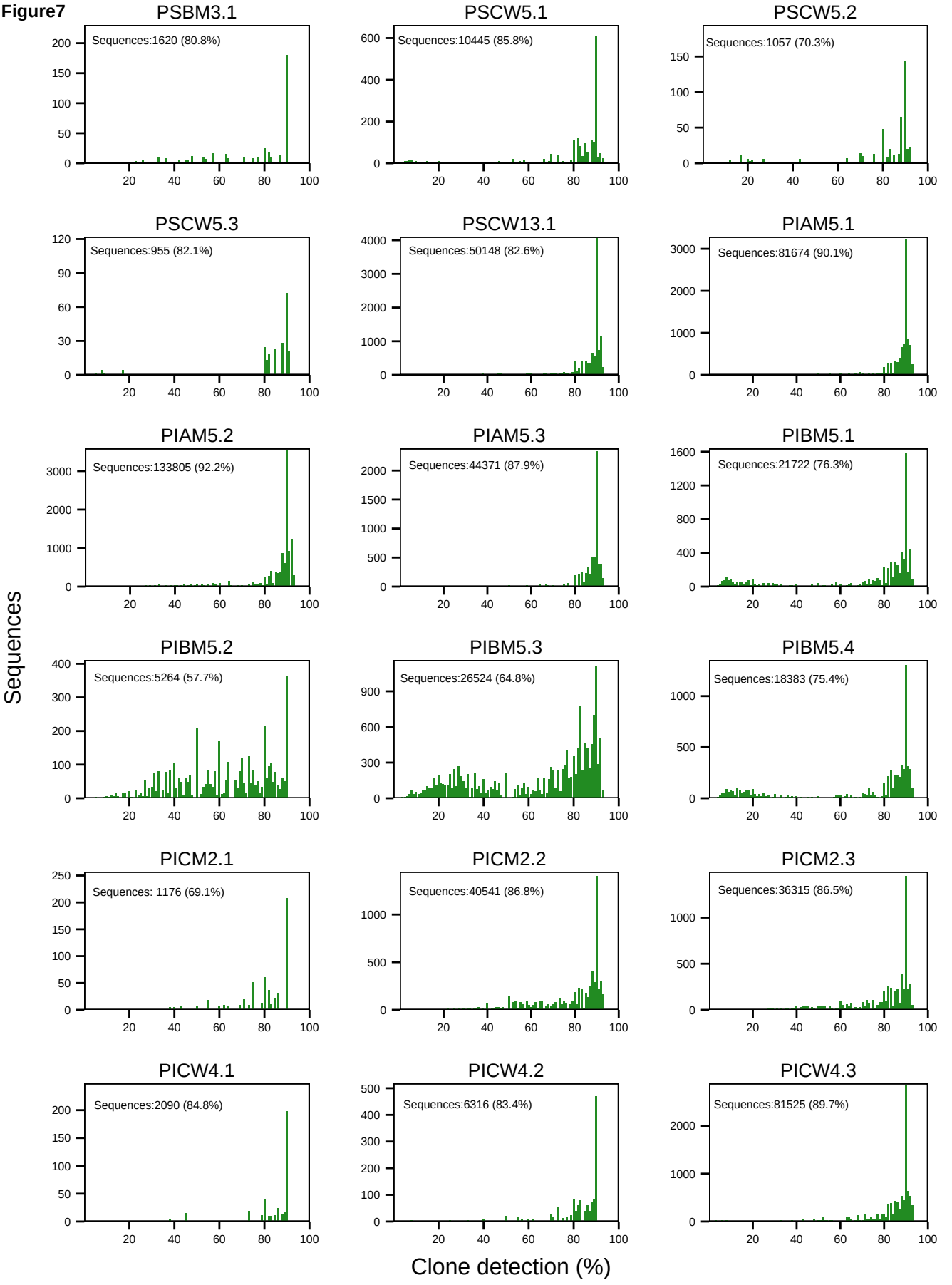
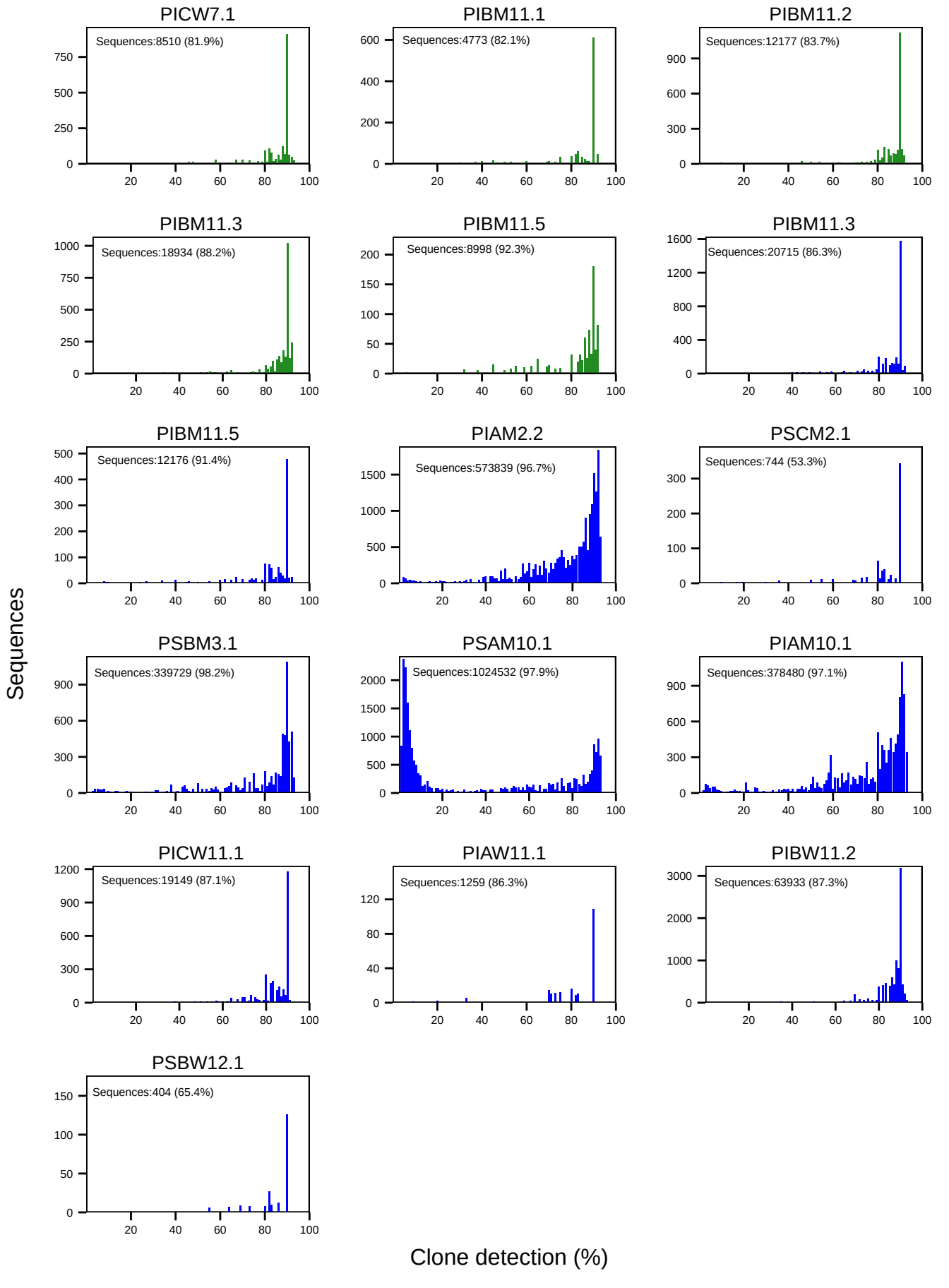


Figure 6

Figure7


PSBM3.1
Sequences:1620 (80.8%)
PSCW5.1
Sequences:10445 (85.8%)
PSCW5.2
Sequences:1057 (70.3%)
PSCW5.3
Sequences:955 (82.1%)
PSCW13.1
Sequences:50148 (82.6%)
PIAM5.1
Sequences:81674 (90.1%)
PIAM5.2
Sequences:133805 (92.2%)
PIAM5.3
Sequences:44371 (87.9%)
PIBM5.1
Sequences:21722 (76.3%)
PIBM5.2
Sequences:5264 (57.7%)
PIBM5.3
Sequences:26524 (64.8%)
PIBM5.4
Sequences:18383 (75.4%)
PICM2.1
Sequences: 1176 (69.1%)
PICM2.2
Sequences:40541 (86.8%)
PICM2.3
Sequences:36315 (86.5%)
PICW4.1
Sequences:2090 (84.8%)
PICW4.2
Sequences:6316 (83.4%)
PICW4.3
Sequences:81525 (89.7%)
Sequences
Clone detection (%)


PICW7.1
Sequences:8510 (81.9%)
PIBM11.1
Sequences:4773 (82.1%)
PIBM11.2
Sequences:12177 (83.7%)
PIBM11.3
Sequences:18934 (88.2%)
PIBM11.5
Sequences:8998 (92.3%)
PIBM11.3
Sequences:20715 (86.3%)
PIBM11.5
Sequences:12176 (91.4%)
PIAM2.2
Sequences:573839 (96.7%)
PSCM2.1
Sequences:744 (53.3%)
PSBM3.1
Sequences:339729 (98.2%)
PSAM10.1
Sequences:1024532 (97.9%)
PIAM10.1
Sequences:378480 (97.1%)
PICW11.1
Sequences:19149 (87.1%)
PIAW11.1
Sequences:1259 (86.3%)
PIBW11.2
Sequences:63933 (87.3%)
PSBW12.1
Sequences:404 (65.4%)
Sequences
Clone detection (%)

Supplementary Information

Methods

Wet lab experimental procedures

Sampling and processing. Sputum was collected within one year on two to four occasions according to the Standard Operating Procedure 530.00 of the CFFT Therapeutics Development Network Coordinating Center [1]. In brief, the subjects performed up to 4 cycles of 3-minutes inhalation of 3% hypertonic saline with the Pari Boy S nebulizer (PARI, Starnberg, Germany). Secretions were mobilized by autogenic drainage in order to ensure representative sampling of the whole lung. Expectorated respiratory secretions were flushed with nitrogen, shock-frozen at -80°C and then stored at -80°C or immediately processed.

Fresh or thawed samples were diluted 1:5 with ice-cold 97.5 % phosphate buffered saline/2.5% mercaptoethanol (v/v) and incubated on ice for 2 h under shaking. The suspension was centrifuged (15 min; 3,800 g; 10°C), the pellet was dried for 10 s, dissolved at 4°C in 10 mL bi-distilled water and then incubated on ice for 15 min on a rocker switch. This cycle of centrifugation, drying and incubation in distilled water was repeated twice. The pellet was dissolved in 1 mL 0.1 RDD-buffer (QIAGEN, Hilden) and incubated in two 0.5 mL aliquots with 60 units DNase I for 90 min at 30°C under shaking (350 rpm). The solutions were combined, diluted with 40 mL DNase buffer and centrifuged (15 min; 3,800 g; 10°C). The pellets were washed three times with 10 mL SE-buffer each by centrifugation, then dissolved in 0.5 mL SE-buffer and pelleted again (10 min, 12,000 g, 10°C). Subsequently genomic DNA was purified according to the 'Hard-to-lyse-Bacteria' protocol with the NucleoSpin Tissue kit (Machery-Nagel, Düren). DNA was stored at 4°C in Tris-EDTA buffer. Yield of double-stranded DNA was determined at the Qubit 1.0 fluorimeter with the Qubit dsDNA BR assay kit (Q32850, Agilent technologies). This protocol was found to be an acceptable compromise to obtain some non-stoichiometric amounts of hard-to-lyse mycobacteria and fungi and not to lose all easy-to-lyse mycoplasms.

DNA library preparation. 0.1 - 1 µg of DNA was sheared in a Covaris S2 system. End repair and size selection to an average of 200 bp fragment size was performed according to standard protocols (Fragment library generation, Life technologies (LT)/Thermo). Standard protocols for the generation of fragment libraries for NGS applications generate a bias for GC-rich sequences, as the ligation of the adaptors becomes inefficient for DNA fragments containing more than 65% GC. To compensate for this constraint, we modified the ligation step of the standard protocol (LT/Thermo). The dA tailing reaction was performed in ¼ of the standard volume with Stratec Taq Polymerase instead of the LT- dA tailing enzyme (DNA 9µl; 5x Buffer(LT) 2.5µl, 10mM dATP 0.25µl, Stratec Taq Polymerase 1.25µl; 30 min; 68°C). The incubation conditions of the subsequent ligation were altered to increase life time and performance of the T4 ligase (dA-tailed reaction mix 13 µl; 5x Buffer(LT) 0.75µl; each adaptor

(LT,1:20 diluted) 0.5 µl; 10mM dNTP 0.3µl; Quick T4 Ligase (NebNext, NEB) 0.8µl; water 0.1µl; 12 h; 12°C; followed by nick translation (20 min; 72°C)). The subsequent purification and amplification (5 cycles) of the generated fragment library was performed according to LT standard protocols.

Sequencing. The binding of the fragment library to beads was performed according to manufacturer's protocols (EZBead System(LT); E120 scale, P2 post enrichment 17%). Sequencing was performed on a SOLiD 5500XL system (LT) with 75 bp read length and implemented Exact call chemistry (LT). The accuracy of sequencing of the instrument was determined independently to be 99.943 %, i.e. a mean of 57 single nucleotide errors are estimated per 100,000 bp of raw sequence.

In silico analyses

Processing of sequences reads. In total we obtained 2.2 billion color space, quality-trimmed and filtered single-end sequences with an average length of 60 bp. More than 77 million reads (3.5% of the total amounts of reads) were non-human (average of 1.25 million reads (74.6 Mbp) per sample). Raw sequencing data were first processed to remove SOLiD barcode sequences and thereafter trimmed to filter out low quality reads. Sequences with at least 40 bases with a quality score above 20 and a minimum length of 45 bp were selected for the analysis. The trimmed reads were aligned against the human reference genome (NCBI build 37/hg19), first using the ultrafast Bowtie2 [2] and thereafter the unaligned reads were processed using the NovoalignCS (http://www.novocraft.com/) short read aligner.

Non-human reads were then checked for low complexity reads. Low-complexity sequences contain repetitions of nucleotides with low or limited information content, e.g. two- or three-letter repeats. These sequences are prone to cause false positive cross-alignments to human and microbes, so they need to be removed. The grade of complexity was estimated by PRINSEQ (http://prinseq.sourceforge.net/) with the DUST [3] method which calculates the frequency distribution of trinucleotides whereby high scores are attributed to mono-, di- or trinucleotide repeats. A stringent threshold of 5 was necessary to eliminate the low complexity reads.

Non-human sequences were corrected with the software SOLiD Accuracy Enhancer Tool (SAET), which increased the number of mapped reads by 40 - 50% in genomes 1 Kbp-200 Mbp in size (http://solidsoftwaretools.com/gf/project/saet/ and http://bcc.bx.psu.edu/download/saet.2.2/) and reduced the error rate by 3 to 5-fold. Reads were grouped by similarity. If a mismatch was found and it was not supported by high quality reads, the software corrected the low quality read having a 'consensus' sequence.

Reference-based taxonomic classification. A local database of complete microbial reference genomes was created (1,800 bacteria, 610 fungi, 5,804 viruses and 5 archaea) downloaded from the National Centre for Biological Information (NCBI, http://www.ncbi.nlm.nih.gov). Draft or

incomplete genomes were not considered because they frequently contain contaminations and implausible sequences.

Non-human corrected reads were aligned to the database using NovoalignCS. The option -r 'All' was used for the identification of bacteria at species level, whereas the more stringent option -r 'None' that searches for unique matches was used for the analysis of viruses, fungi and bacteria at the strain level. Reads aligned to multiple bacterial genomes were interrogated with an in-house Perl script whether they could be reassigned to the species level.

Removal of mobile genetic elements. First, a single-sample t-test method was applied to calculate the mean distances among the reads aligned to a specific reference genome. A cutoff p-value of $p < 0.01$ removed most sequences clustered to a specific region. However, in cases of numerous genomic islands in a bacterial genome, the distribution of distances of genome map positions between pairs of sequence reads was interrogated whether it followed a Gaussian distribution centered around '0.25 x genome size' [4].

Normalization. The remaining microbial reads were then normalized by GC content and genome length. The SOLiD technology has a pronounced GC bias in GC-rich regions [5] which affects the quantification of microbial genome abundances. Based on sequencing of a set of bacterial strains of 30% to 71% GC-content on the SOLiD instrument, an empirical algorithm was developed that normalizes each read by its normalized coverage coefficient (dependent variable), based on the GC content of the read and the GC content of the genome to which it has aligned (independent variables) (Chouvarine et al., unpublished). The GC-corrected reads were then normalized by genome length and reported as counts per Mb of reference. Finally, bacterial abundance was normalized to bacterial DNA per human cell present in the metagenomic sample.

Unaligned sequences (from 0.5% to 28.9 % of total amount of sequences per sample) were queried by blastn against the NCBI nt database (downloaded in January 2014) to improve the recovery rate of rare species or incomplete genomes not present in our database. Default values were selected to take the best hit for each sequence match.

Principal component analysis. We performed principal component analysis of bacterial abundances on the genus level of the samples divided into pancreatic sufficient (PS) and pancreatic insufficient (PI) groups. This analysis was performed using two different methods. In the first method the relative bacterial abundances were created by applying the decostand (data,"total") method from the vegan package for R on the abundance count data followed by application of the prcomp function in R for standard PCA. In the second method we used absolute bacterial abundances per human cell to perform the same analysis. In both cases the bacterial data were normalized for GC bias and genome length of the bacterial reference genomes.

***P. aeruginosa* and *S. aureus* clone analysis.** *P. aeruginosa* and *S. aureus* sequences were aligned with NovoalignCS against the *P. aeruginosa* PAO1 [6] and *S. aureus* Newman [7] reference genomes. Single nucleotide polymorphisms (SNPs) were extracted using Samtools [8]. To determine the number of clones and their relative abundance, the ratio of mismatches to matches was counted for each SNP at genome positions covered by more than 10 reads whereby the match represents the nucleotide of the most frequent clone. Each clone with n_i SNPs and a relative abundance p ($0 < p \leq 1$) in the clonal population will show a hypergeometric distribution of hits at j genome positions. Considering the accuracy of SOLiD technology of 99.94%, only clones with a relative abundance of at least 0.1% will show reliable signals in a data set of at least 10,000 species-specific reads. For *P. aeruginosa* clone type identification, 12 SNPs in seven loci of the core genome were queried that had previously been selected for a multi-marker genotyping device [9]. *S. aureus* clone types were identified by sequence type. Sequence types were downloaded from http://saureus.mlst.net/ and the experimental reads were aligned with NovoalignCS against them. The analysis of *P. aeruginosa* and *S. aureus* clones was performed on 25 samples from 10 subjects and 16 samples from 13 individuals, respectively.

Antimicrobial resistance genes identification. Bacterial sequences were aligned to 'The Comprehensive Antibiotic Resistance Database' (CARD) [10] to define genetic carriage of resistance profiles in the cystic fibrosis lungs.

Uniquely aligned reads carrying a maximum number of 3 SNPs were selected for analysis. *P. aeruginosa* and *S. aureus* sequences aligned against CARD were extracted and aligned (following the same procedure described previously) against the *P.aeruginosa* PAO1 and *S. aureus* Newman reference genomes, respectively. Samtools and SnpEff [11] were used to extract and categorize the effects of the genetic variants on the coding DNA sequences. We detected 132 SNPs present in the aligned *P. aeruginosa* reads, 20 (12.5%) of which were non-synonymous SNPs. Five SNPs were qualified as rare or *de novo* mutations present in less than 20% of the aligned sequences. The *S. aureus* genes contained 221 SNPs and 2 indels of which 30 were non-synonymous SNPs (13 of them rare or *de novo* mutations).

Statistical and phylogenetic analysis. R software was used to perform all statistical analysis. The program MetaPhlAn2 [12, 13] was used for taxonomic classification of normalized sequence data and for the construction of heatmaps of the most abundant species. Trees of life were constructed with the tool GraPhlAn [14] (Graphical Phylogenetic Analysis).

References

1. CFFT Therapeutics Development Network. Sputum induction using the Nouvag nebulizer with medication cup. 2010.
2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012; **9:** 357–359.
3. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low complexity DNA sequences. *J Comput Biol* 2006; **13:** 1028-1040.
4. Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, Paetow M, Siekmann B, Wieding-Drewes M, Wienhöfer M, Wolf S, Tümmler B, Ahlers V,Sprengel F. Genometa--a fast and accurate classifier for short metagenomics shotgun reads. *PLoS One* 2012; **7**: e41224.
5. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P, Pfister S, Wolf S, Brors B, Eils R. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One* 2013; **8**: e66621.
6. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E,Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE, Lory S, Olson MV. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 2000; **406**: 959-964.
7. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *J Bacteriol* 2008; **190**: 300-310.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078-2079.
9. Wiehlmann L, Cramer N, Tümmler B. Habitat-associated skew of clone abundance in the *Pseudomonas aeruginosa* population. *Environ Microbiol Rep* 2015 Sep 30. doi: 10.1111/1758-2229.12340.
10. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I,Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013; **57**: 3348-3357.
11. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6**: 80-92.
12. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *Peer J* 2015; **3:** e1029.
13. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012; **9**: 811-814.
14. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015; 12: 902-903.

Figure Legends

Figure S1. Rank number differences between the frequency of detection and the relative abundance of taxa in the whole data set of metagenomes of CF sputa. All detected taxa were sorted by rank numbers for the total number of reads assigned to the respective taxon and its detection rate in the samples. The figure displays the difference of rank numbers between abundance and frequency of detection for the top 95% of species belonging to the actinobacteria, bacteroidetes, firmicutes, fusobacteria or proteobacteria, respectively. Rank number differences are shown for samples collected from PI (green triangle), PS (blue square) and all patients with CF (orange circle).

Figure S2. Heatmaps of the relatedness of sputum metagenomes of PI (**a**) and PS (**b**) patients based on the abundance of the most frequent bacterial species.

Figure S3. Association of the detection rate of microbial species with the total number of assigned microbial reads in the metagenome sample. The species composition of the individual metagenomes is shown whereby the color of a dot visualizes the detection rate of the respective species in all specimens.

Table Captions

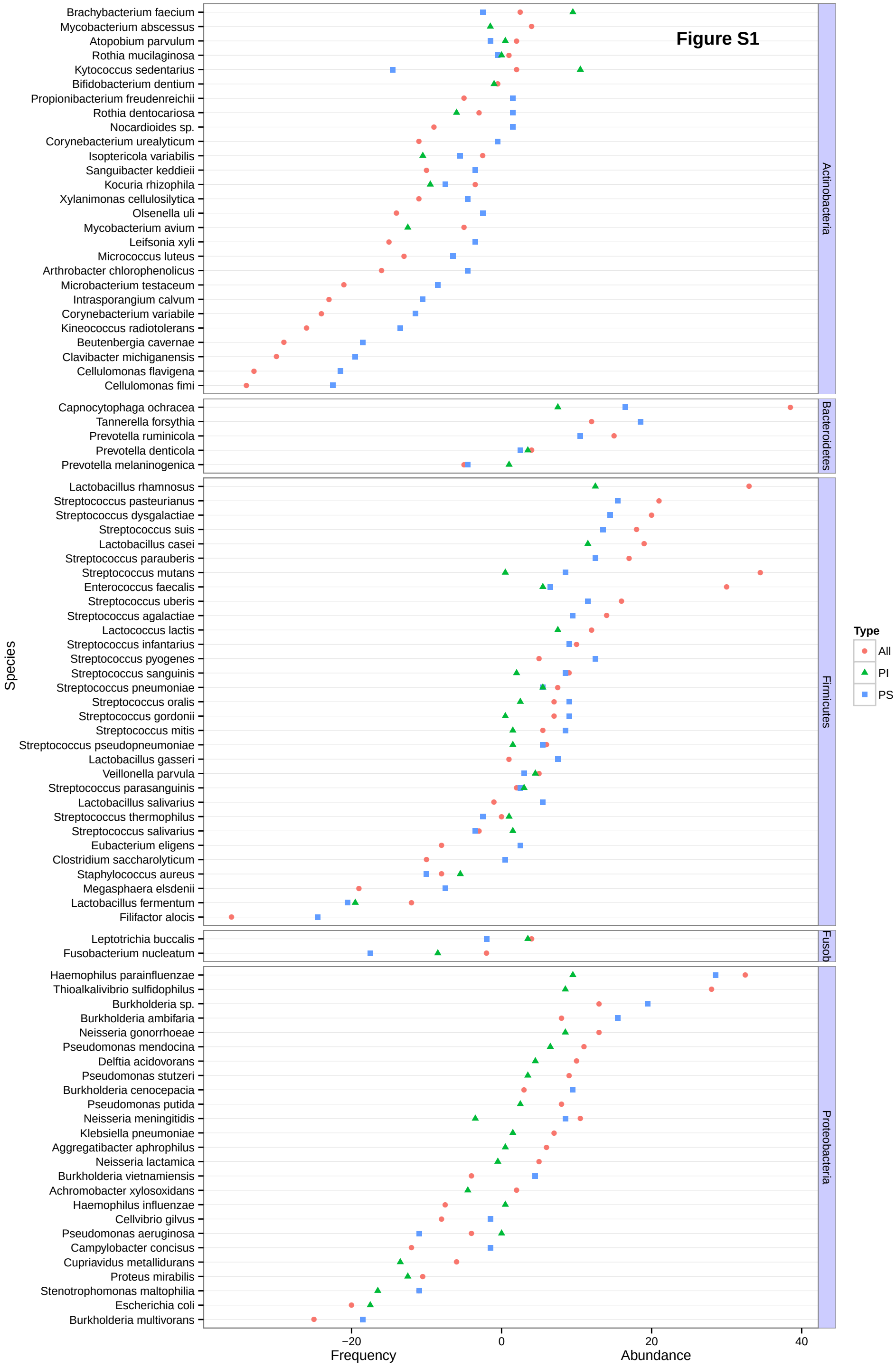
Table S1. Reads of DNA viruses, bacteria, molds and fungi detected at the species level in the individual sputum metagenomes collected from patients with cystic fibrosis.

Table S2. A. Number of species (DNA viruses, bacteria, fungi) detected in sputa collected from pancreatic exocrine sufficient (PS) or insufficient (PI) children (group A, 8-13 years), adolescents and young adults (group B, 18-23 years) and adults (group C, > 28 years) with cystic fibrosis. **B.** Relative abundance in per cent of DNA viruses, bacteria and fungi in the individual CF sputa. **C.** Proportion of anaerobes among the bacteria in the sputum metagenomes. **D.** Number of sequence reads of bacteriophages and their respective bacterial hosts.

Table S3. Normalized abundance and detection rates of microbial species in the sputum metagenomes differentiated by bacteria, DNA viruses and eukaryotic microbes (molds and fungi).

Table S4. Clonal diversity of *S. aureus* and *P. aeruginosa* populations in CF sputum. Number of reads that at SNP positions were divergent (‘mismatch’) from the nucleotide of the most prevalent clone (‘match’) sorted in 1%-intervals of the mismatch/match ratio of SNP-encoding reads. Only SNPs were considered that were covered by more than ten sequence reads.

Figure S1


Species
Brachybacterium faecium
Mycobacterium abscessus
Atopobium parvulum
Rothia mucilaginosa
Kytococcus sedentarius
Bifidobacterium dentium
Propionibacterium freudenreichii
Rothia dentocariosa
Nocardioides sp.
Corynebacterium urealyticum
Isoptericola variabilis
Sanguibacter keddieii
Kocuria rhizophila
Xylanimonas cellulosilytica
Olsenella uli
Mycobacterium avium
Leifsonia xyli
Micrococcus luteus
Arthrobacter chlorophenolicus
Microbacterium testaceum
Intrasporangium calvum
Corynebacterium variabile
Kineococcus radiotolerans
Beutenbergia cavernae
Clavibacter michiganensis
Cellulomonas flavigena
Cellulomonas fimi
Actinobacteria
Capnocytophaga ochracea
Tannerella forsythia
Prevotella ruminicola
Prevotella denticola
Prevotella melaninogenica
Bacteroidetes
Lactobacillus rhamnosus
Streptococcus pasteurianus
Streptococcus dysgalactiae
Streptococcus suis
Lactobacillus casei
Streptococcus parauberis
Streptococcus mutans
Enterococcus faecalis
Streptococcus uberis
Streptococcus agalactiae
Lactococcus lactis
Streptococcus infantarius
Streptococcus pyogenes
Streptococcus sanguinis
Streptococcus pneumoniae
Streptococcus oralis
Streptococcus gordonii
Streptococcus mitis
Streptococcus pseudopneumoniae
Lactobacillus gasseri
Veillonella parvula
Streptococcus parasanguinis
Lactobacillus salivarius
Streptococcus thermophilus
Streptococcus salivarius
Eubacterium eligens
Clostridium saccharolyticum
Staphylococcus aureus
Megasphaera elsdenii
Lactobacillus fermentum
Filifactor alocis
Firmicutes
Leptotrichia buccalis
Fusobacterium nucleatum
Fusob
Haemophilus parainfluenzae
Thioalkalivibrio sulfidophilus
Burkholderia sp.
Burkholderia ambifaria
Neisseria gonorrhoeae
Pseudomonas mendocina
Delftia acidovorans
Pseudomonas stutzeri
Burkholderia cenocepacia
Pseudomonas putida
Neisseria meningitidis
Klebsiella pneumoniae
Aggregatibacter aphrophilus
Neisseria lactamica
Burkholderia vietnamiensis
Achromobacter xylosoxidans
Haemophilus influenzae
Cellvibrio gilvus
Pseudomonas aeruginosa
Campylobacter concisus
Cupriavidus metallidurans
Proteus mirabilis
Stenotrophomonas maltophilia
Escherichia coli
Burkholderia multivorans
Proteobacteria
−20
0
20
40
Frequency
Abundance
Type
All
PI
PS

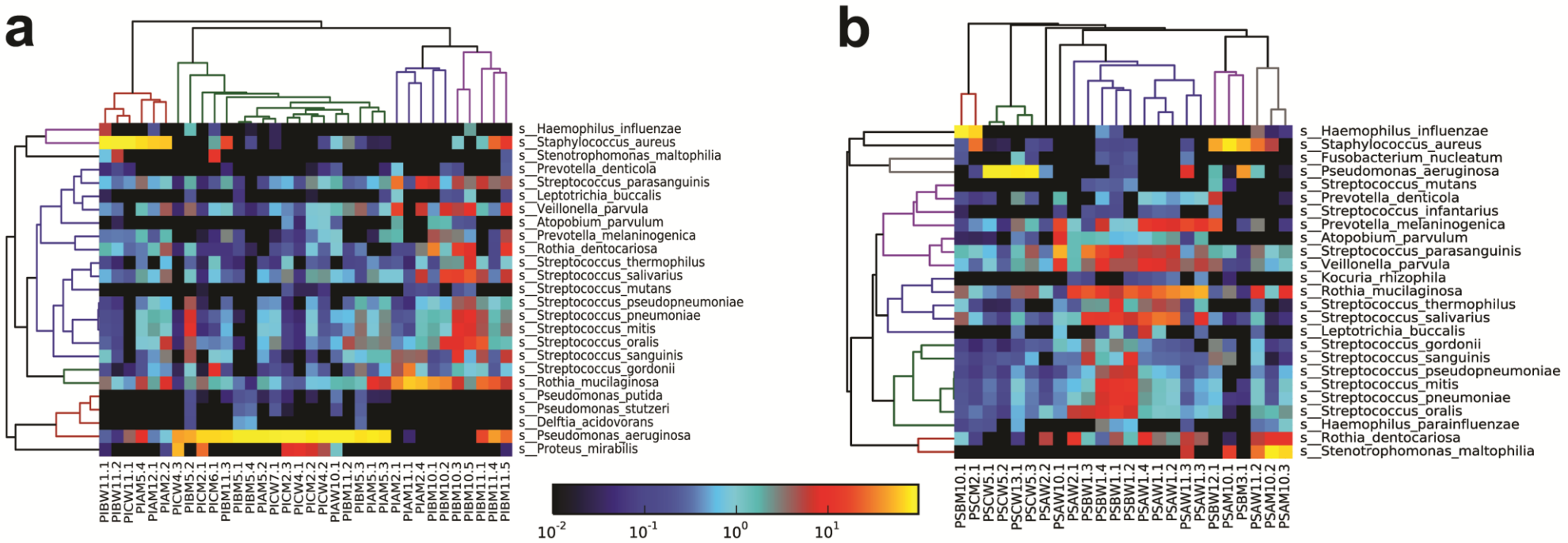

a
s__Haemophilus_influenzae
s__Staphylococcus_aureus
s__Stenotrophomonas_maltophilia
s__Prevotella_denticola
s__Streptococcus_parasanguinis
s__Leptotrichia_buccalis
s__Veillonella_parvula
s__Atopobium_parvulum
s__Prevotella_melaninogenica
s__Rothia_dentocariosa
s__Streptococcus_thermophilus
s__Streptococcus_salivarius
s__Streptococcus_mutans
s__Streptococcus_pseudopneumoniae
s__Streptococcus_pneumoniae
s__Streptococcus_mitis
s__Streptococcus_oralis
s__Streptococcus_sanguinis
s__Streptococcus_gordonii
s__Rothia_mucilaginosa
s__Pseudomonas_putida
s__Pseudomonas_stutzeri
s__Delftia_acidovorans
s__Pseudomonas_aeruginosa
s__Proteus_mirabilis
PIBW11.1
PIBW11.2
PICW11.1
PIAM5.4
PIAM12.1
PIAM2.2
PICW4.3
PIBM5.2
PICM2.1
PICM6.1
PIBM11.3
PIBM5.1
PIBM5.4
PIAM5.2
PICW7.1
PICM2.3
PICW4.1
PICM2.2
PICW4.2
PIAW10.1
PIBM11.2
PIBM5.3
PIAM5.1
PIAM5.3
PIAM2.1
PIAM11.1
PIAM2.4
PIBM10.1
PIBM10.2
PIBM10.3
PIBM10.5
PIBM11.1
PIBM11.4
PIBM11.5
10^{-2}
10^{-1}
10^{0}
10^{1}
b
s__Haemophilus_influenzae
s__Staphylococcus_aureus
s__Fusobacterium_nucleatum
s__Pseudomonas_aeruginosa
s__Streptococcus_mutans
s__Prevotella_denticola
s__Streptococcus_infantarius
s__Prevotella_melaninogenica
s__Atopobium_parvulum
s__Streptococcus_parasanguinis
s__Veillonella_parvula
s__Kocuria_rhizophila
s__Rothia_mucilaginosa
s__Streptococcus_thermophilus
s__Streptococcus_salivarius
s__Leptotrichia_buccalis
s__Streptococcus_gordonii
s__Streptococcus_sanguinis
s__Streptococcus_pseudopneumoniae
s__Streptococcus_mitis
s__Streptococcus_pneumoniae
s__Streptococcus_oralis
s__Haemophilus_parainfluenzae
s__Rothia_dentocariosa
s__Stenotrophomonas_maltophilia
PSBM10.1
PSCM2.1
PSCW5.1
PSCW5.2
PSCW13.1
PSCW5.3
PSAW2.2
PSAW10.1
PSAW2.1
PSBW1.3
PSBW1.4
PSBW1.1
PSBW1.2
PSAW1.4
PSAW1.1
PSAW1.2
PSAW11.3
PSAW1.3
PSBW12.1
PSAM10.1
PSBM3.1
PSAW11.2
PSAM10.2
PSAM10.3

Figure S2

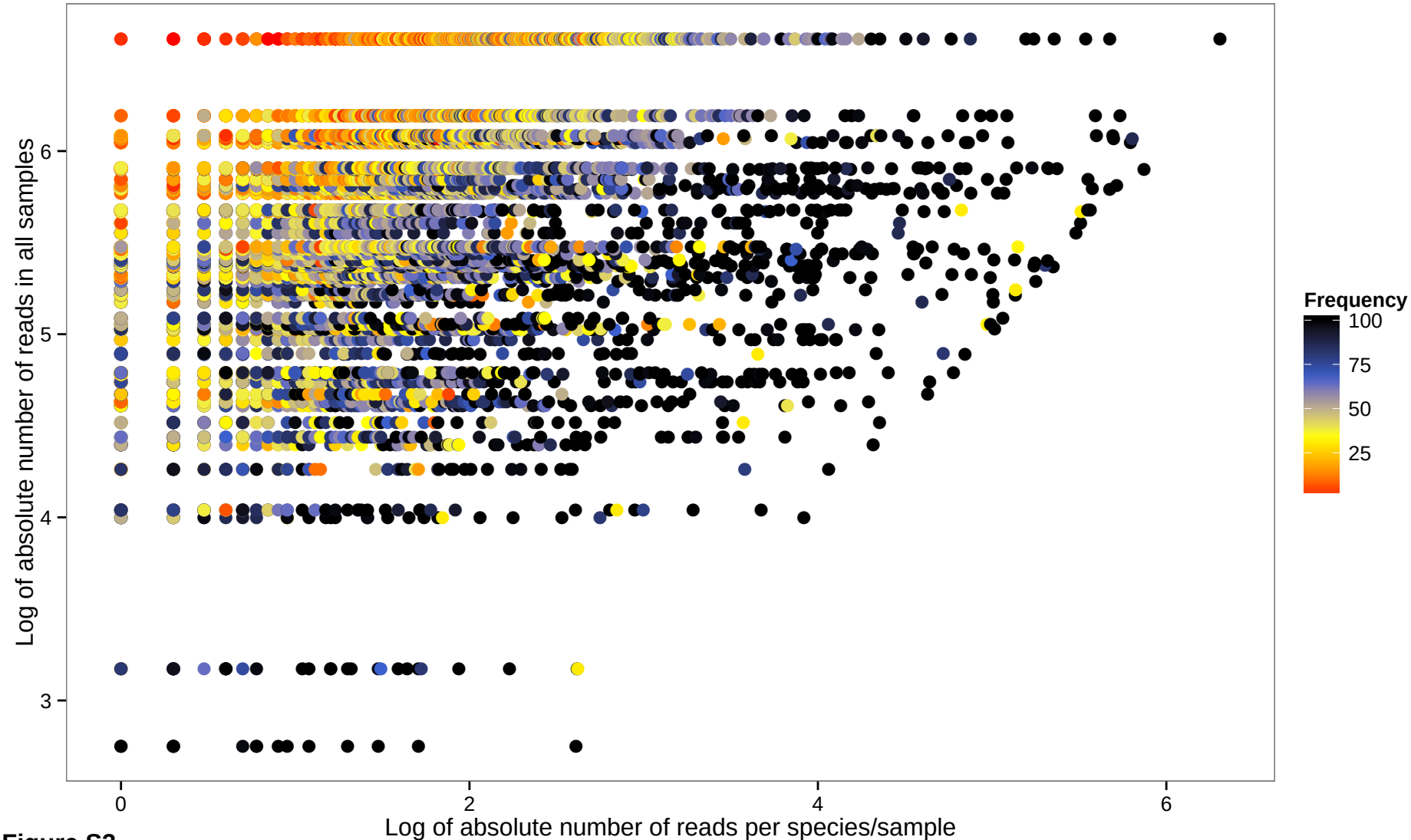

Log of absolute number of reads in all samples
6
5
4
3
0
2
4
6
Log of absolute number of reads per species/sample
Frequency
100
75
50
25

Figure S3

Chapter 8

Conclusion and Future Perspectives

In this chapter, I reiterate the research objectives of this thesis and describe the major findings, including their implications, limitations and future perspectives. I also give a short comment on my personal view on the future of metagenomic studies and bioinformatics as well as potential roadblocks that may need to be addressed to continue improving bioinformatics and microbial genetics research. Finally, I point out the contributions of this work to the cystic fibrosis community and recommendations for whole metagenome sequencing project based on our findings.

8.1 Thesis Research Objective: Major Findings, Implications, Limitations and Future Perspectives

Next-generation sequencing technologies (NGS) have had an incredible impact in the field of genomic research increasing our knowledge of microbial population and their implications in human health. To bypass the limitations of culturable bacteria techniques[93], 16S rDNA and whole-genome shotgun sequencing[94-96] have become popular culture-independent methods used to analyze microbial samples and explore the population structures, genetic diversity and interactions of microbial communities directly from their habitats.

For the first time, we can get reliable descriptions of the microbes which live in the human body (human microbiome) and their impact on different diseases like type 2 diabetes and obesity[97-98]. Furthermore, we can start to understand how microbes adapt to their environments and evolve creating serious problems to human health such as antibiotic resistance. These are just a few examples where metagenomic studies open up fundamental applications in medical research, where personalized healthcare is the most promising area.

However, metagenomic analyses present several challenges in order to perform a meaningful and useful study. Taxonomic classification, gene prediction, assembly or biodiversity estimation are some examples of these difficulties. Also, the grade of complexity present in the community

and the large number of sequence data can complicate the bioinformatic analysis. An understanding of the biases and errors in the sequencing data and a continued improvement of bioinformatics tools is essential in order to achieve an accurate research project.

This thesis presents projects which concern the cystic fibrosis lower airways microbiome, where the main focus is on the whole metagenome sequencing analysis. Knowing that pathogenic and non-pathogenic bacteria can belong to the same species, a novel algorithm was developed to calculate bacterial relatedness based on the detection of bacterial haplotypes. Furthermore, a new model of SOLiD sequencing normalization and the application of the bacterial recombination approach on the two main cystic fibrosis pathogens (*P. aeruginosa* and *S.aureus*) are presented.

8.1.1. The cystic fibrosis lower airways microbial metagenome

The work was centered on the analysis of cystic fibrosis metagenome sputum samples of the lower airways using whole-genome shotgun sequencing. The identification of frequencies and abundances of viruses, bacteria and fungi in cystic fibrosis individuals becomes crucial to understand the evolution of the sickness. Previous analyses based on the 16S rDNA sequencing already have provided an overview of the taxa present in the lower airways of individuals with cystic fibrosis[99-101]. However, as have been mentioned in Chapter 2, this approach has some limitations.

In order to gain a solid and better understanding of the cystic fibrosis microbial communities I conducted to date the largest and deepest cystic fibrosis metagenome analysis covering all different age groups and states of the disease, to comprehensively characterize the microorganisms present in the community. For the first time, an exhaustive analysis of respiratory secretions of exocrine pancreas sufficient (PS) and exocrine pancreas insufficient (PI) subjects with cystic fibrosis of different age groups (children, adolescents and adults) has been performed. My study confirmed the presence of a large and diverse repertoire of microbial taxa in the CF lower airway where bacteria typically made up more than 99% of the microbial community, while viruses and fungi were present in less than 1%. An individual signature of multiple species present in low abundance and few disease-associated pathogens (such as *P. aeruginosa* and *S. aureus*) in high abundance constitute the polymicrobial community. Furthermore, I was able to estimate the proportion of anaerobes in the microbial metagenome which decreases with age in our cohort of CF patients. The decline of the anaerobes is replaced by the leading pathogen *P. aeruginosa* as the severity of the disease increases with aging.
Armed with this knowledge, a detailed overview of the relative abundance of all detected species is now presented. The dominant taxa in the metagenome of the whole cohort were the bacterial species that belong to the groups streptococci, staphylococci, pseudomonads as well

as Haemophilus sp., Burkholderia sp. and Stenotrophomonas sp.. The leading pathogen *P. aeruginosa* was identified in all metagenomic data sets from PI patients although six of them had been classified as *P. aeruginosa* negative by culture-dependent analysis. This clinically important finding suggests the ubiquitous presence of *P. aeruginosa* in respiratory secretions of PI CF patients although its relative abundance was just 0.02% in *P. aeruginosa* – negative patients. The viral community consisted primarily of phages, a few human pathogens, primarily herpes virus and adenovirus, and rare cases of viruses infecting non-mammalian eukaryotic hosts. Lastly, the mycobiome community was dominated by Aspergillus species and Saccharomycetes including Candida sp..

In addition, the clonal composition of the populations of the leading pathogens *S.aureus* and *P.aeruginosa* was identified from the frequency distribution of SNPs in the metagenomes of our CF cohort. The *S. aureus* and *P. aeruginosa* populations were found to be composed of one major and numerous minor clone types. The rare clones constitute a low copy genetic resource which could rapidly expand as a response to habitat alterations such as antimicrobial chemotherapy or invasion of novel microbes. Conversely, the low coverage of sequences repeatedly prevented the identification of the genotype of the dominant strain within the frame of established typing schemes. Nevertheless, four of ten analyzed *P. aeruginosa* strains were assigned to ubiquitous clones of the global *P. aeruginosa* population and two pairs out of 13 *S. aureus* strains were identified to belong to the common sequence type ST7 and the pandemic MRSA lineage ST22, respectively.

Finally, I performed the analysis of antibiotic resistance genes detection in *S. aureus* and *P. aeruginosa*. Uncommon non-synonymous nucleotide substitutions were present in *mexS*, *mexF*, *mexI* and *aph P. aeruginosa* genes which encode drug-inactivating enzymes or multidrug efflux pumps. In *S. aureus* mutations were detected in *gyrA*, *gyrB*, *rpoB*, *rpoC*, *tufA*, *parE* and *parC.*

These results provide significant and novel knowledge for the cystic fibrosis research community.

8.1.2. Detection of recombination in bacterial genomes by haplotype construction.

The second objective of my PhD was the study of bacterial recombination in the two dominant pathogens of cystic fibrosis, *P. aeruginosa* and *S. aureus*. Recombination is a key process in bacterial evolution, therefore it is crucial that we have bioinformatic tools at hand that are able to precisely detect its occurrence and interpret its effects within phylogenetic relationships.

In analogy to the diploid genomes, in this approach the term haplotype is defined as the number of syntenic SNPs in paired comparisons of bacterial genomes. I developed a novel algorithm to

analyze homologous recombination in the bacterial core genome of bacteria based on haplotype reconstruction. To identify the haplotype blocks, each strain was aligned against the reference genome, and a first matrix was created which contained all SNPs detected for each of these strains. These SNPs were then ordered by genome position. Thereafter, pairwise comparisons were performed to construct a second matrix. This second matrix contained the number of consecutive shared SNPs at each SNP position for each pair. This approach had to rely on the estimation of the minimum length of SNPs shared between genomes, since the start and stop positions of a haplotype are not able to be detected. With this fast and simple approach the identical fragments present between two strains, are detected. The structure and homology present in the bacterial community are calculated from the frequency and size distribution of haplotypes. Therefore, related strains will share longer haplotypes than unrelated strains.

To avoid making assumptions on the complexity and structure of the population, I applied my algorithm to three different projects with real data sets:

a) Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment.

In this study, the 15 most frequent clonal complexes in the *P. aeruginosa* population and the 5 most common clones from the environment were analyzed.
Performing the reconstruction of haplotypes, 192,443 quality-controlled SNPs shared at least in two strains, were identified in the 210 paired comparisons whereby PAO1 was the reference genome. These SNPs were exploited to construct a total of 3,779,224 SNP syntenies (‘haplotypes’) with lengths between 2 and 2,348 consecutive-shared SNPs. The median physical length of paired conserved sequence was calculated to be 207 base pairs suggesting an unrestricted gene flow between clonal complexes by recombination. The two most related strains (1BAE and 3C2A) shared 70% of the longest haplotypes (≥ 20kbp) implying that they have emerged from a common ancestor.

b) Intraclonal genome diversity of the major Pseudomonas aeruginosa clones C and PA14.

A second project where I applied the reconstruction of haplotypes with the novel algorithm was to study the conservation and relatedness of the two major clones of *P. aeruginosa* clones C and PA14. For the 58 clone C and 42 PA14 isolates, 1653 and 861 pairwise combinations were conducted, respectively. The analysis showed that haplotypes are 1000-fold longer within a clone than among unrelated clones, indicating that the chromosomal frame is conserved among members of a clonal complex.

c) Reconstruction of haplotypes in S. aureus.

Forty-one genome sequences of *S. aureus* were taken for the haplotype reconstruction approach. In this case, the strain "Newman" was the reference. 136,258 SNPs were detected to build up a total number of 8,704,567 haplotypes. The largest haplotype consisted of 2,450,522 nucleotides. A detailed analysis of the *S. aureus* ST5 revealed a median length of paired conserved sequences of 33 syntenic SNPs, which represent a median physical length of 4.2 kbp.

I conclude from the application of this novel algorithm that paired whole genome comparisons of haplotype length allow an unbiased analysis of the population structure of being more clonal or more sexual. Haplotype length calculated as the physical length (nucleotide length) rather than the number of syntenic SNPs (consecutive shared SNPs) provides a better evaluation of relatedness of strains.

8.1.3. Filtration and normalization of sequencing read data in whole-metagenome shotgun samples

Another focus of my thesis was to implement an automatic pipeline for the metagenome analysis of SOLiD technology sequences due to the small number of tools available for this technology. The estimation of bacterial abundances by whole-metagenome shotgun (WMS) sequencing is proportional to the counts of reads mapped to a reference. Based on the biased results obtained from the SOLiD sequencing of different bacterial species with large variation in their GC genome content, a key step in the achievement of an accurate analysis was the implementation of the normalization based on the GC content of the species. A second challenge on the WMS analysis was to identify sequences which belong to genomic islands and can distort the results. Finally, a normalization based on the length of the reference genome must be considered.

A new model which covers all these obstacles by filtration and normalization procedures was developed, thus leading to more accurate estimation of bacterial abundances in a metagenome.

8.2 The Future of Metagenomics and Bioinformatics

The fast advances in sequencing technologies are moving towards the transformation of research areas like metagenomics, human genomics and medical diagnostics. Metagenomics, the cornerstone of this thesis, has bypassed the need of isolation or cultivation of microorganisms present in culture-dependent methods. However, considerable bottlenecks and bioinformatic challenges need to be addressed.

Metagenomics and medical diagnosis.

As one of the most promising applications in clinical microbiology, metagenomics has the potential to revolutionize how pathogens are detected and optimize treatments strategies[102-103]. Profiling studies of the gut microbiome have already shown that the microbiota rather than a single pathogen play an important role in regulating inflammatory and metabolic conditions in determined diseases, such as colon cancer[104-105] and inflammatory bowel disease[106-107]. Therefore, metagenomics offers a better understanding of the entire microbial community and the mechanisms involved in the relationship with the host. In addition, the identification of individual organisms which could confer virulence and antibiotic resistance properties is also possible following this approach.

Third generation sequencing technologies as well as the development of new bioinformatic tools bear the potential to reduce the turnover time of infectious disease diagnostics in the real life medical setting.

Bioinformatic and sequencing challenges

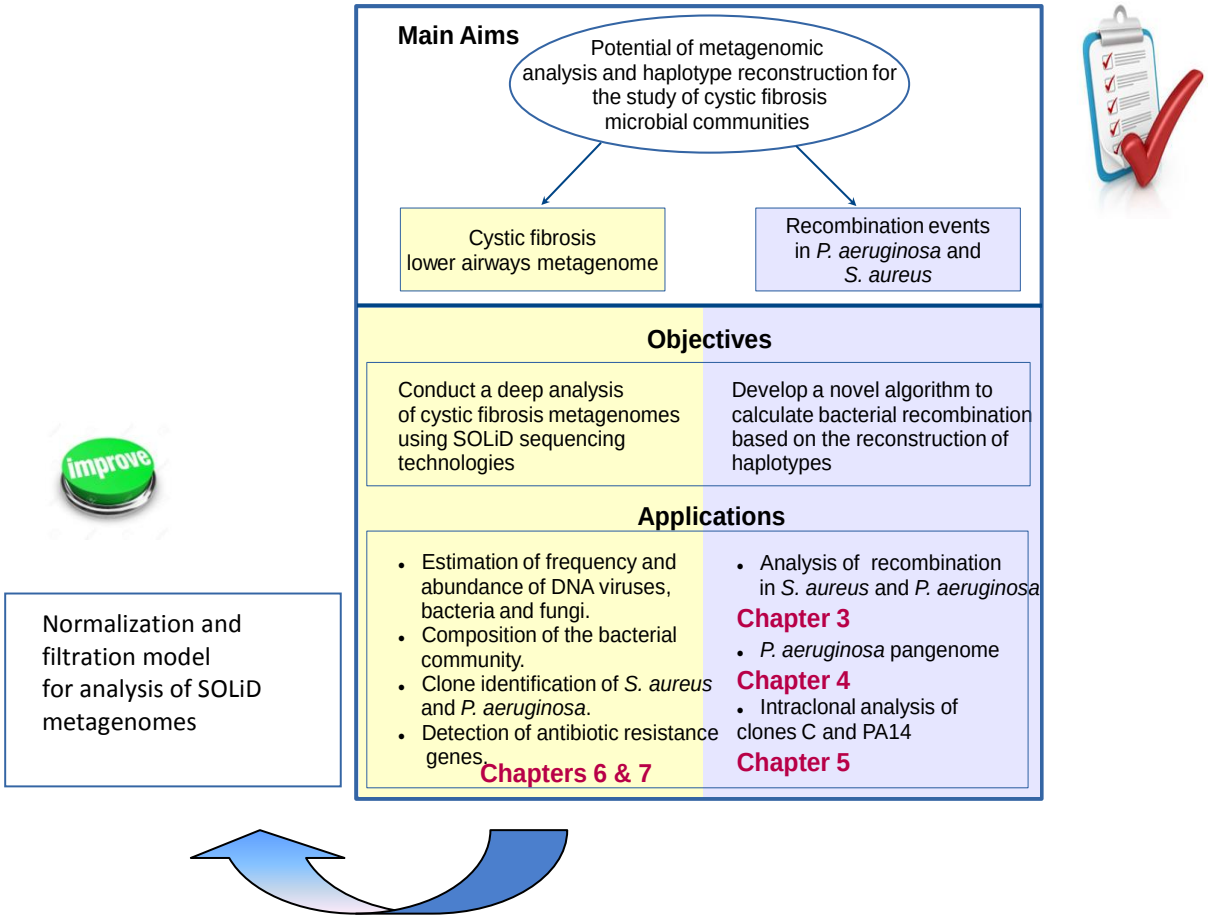
At present, one of the challenges that metagenomics is facing are the computational limitations. These limitations do not just arise from data processing and dearth of appropriate programs and tools. Processing and analysis of big data generated by the new sequencing technologies will be cost-intensive which probably most of the laboratories will not be able to afford[108]. However, cloud computing is a possible solution to overcome these problems.

Other challenges come directly from the accuracy of data and its interpretation as well as from the transformation of the discoveries into medical practice.

Errors generated during sequencing or deposited in databases[109] impede the discovery of novel variants. Detection of SNPs, insertion-deletion variants (indels), structural variants (SVs) and copy number variants (CNVs) is still error-prone. Assemblies of short metagenomic sequences as well as *de novo* assemblies remain a further big obstacle in the analysis of genomic data. Lastly, the methods to understand the functional relationship between associated variants and phenotypic traits must be improved. Therefore, improvements and developments of algorithms and quality control measures are crucial and must be taken into consideration to address all different challenges we are confronting.

8.3 Conclusion

In conclusion, this thesis describes a large and comprehensive metagenome study of the lower airways cystic fibrosis microbiome. It addresses important topics like the identification of viruses or bacterial clones in the metagenome, and the relatedness between microbial species and disease development. As well, an estimation of the recombination rates of the major cystic fibrosis pathogens based on the reconstruction of haplotypes is performed using a novel algorithm for an unbiased analysis. I can confirm that all objectives have been satisfactorily achieved, and even a new model of metagenomic analysis has been developed which was not one of the primary objectives of the dissertation.



Future large-scale studies will hopefully integrate metagenomics with metatranscriptomics, metaproteomics and metabolomics to get an in-depth understanding of the functional dynamics of microbial communities in cystic fibrosis.

Chapter 9

References

1. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667):66-74.
2. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 95(12):6578-83. Review.
3. Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH (2013) Going back to the roots: the microbial ecology of the rhizosphere. Nat Rev Microbiol 11(11):789-99.
4. Wang WL, Xu SY, Ren ZG, Tao L, Jiang JW, Zheng SS (2015) Application of metagenomics in the human gut microbiome. World J Gastroenterol 21(3):803-14.
5. Gibson MK, Pesesky MW, Dantas G (2014) The yin and yang of bacterial resilience in the human gut microbiota. J Mol Biol 426(23):3866-76. Review.
6. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker D, Raes J, Hansen T; MetaHIT consortium, Bork P, Wang J, Ehrlich SD, Pedersen O (2013) Richness of human gut microbiome correlates with metabolic markers. Nature 500(7464):541-6.
7. Keim NL, Martin RJ (2014) Dietary whole grain–microbiota interactions: insights into mechanisms for human health. Adv Nutr 5(5):556-7.
8. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature 457(7228):480-4.
9. Power SE, O'Toole PW, Stanton C, Ross RP, Fitzgerald GF (2014) Intestinal microbiota, diet and health. Br J Nutr 111(3):387-402. Review.
10. Human Microbiome Project Consortium (2012) A framework for human microbiome research. Nature 486(7402):2015-21.
11. Sanschagrin S, Yergeau E (2014) Next-generation sequencing of 16S ribosomal RNA gene amplicons. J Vis Exp(90).
12. Coburn B, Wang PW, Diaz Caballero J, Clark ST, Brahma V, Donaldson S, Zhang Y, Surendra A, Gong Y, Elizabeth Tullis D, Yau YC, Waters VJ, Hwang DM, Guttman DS (2015) Lung microbiota across age and disease stage in cystic fibrosis. Sci Rep 5:10241.
13. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI (2012) Human gut microbiome viewed across age and geography. Nature 486(7402):222-7.
14. Oh J, Byrd AL, Deming C, Conlan S; NISC Comparative Sequencing Program, Kong HH, Segre JA (2014) Biogeography and individuality shape function in the human skin metagenome. Nature 514(7520):59-64.
15. Ward TL, Hosid S, Ioshikhes I, Altosaar I (2013) Human milk metagenome: a functional capacity analysis. BMC Microbiol 13:116.

16. Be NA, Thissen JB, Fofanov VY, Allen JE, Rojas M, Golovko G, Fofanov Y, Koshinsky H, Jaing CJ (2015) Metagenomic analysis of the airborne environment in urban spaces. Microb Ecol 69(2):346-55.

17. Hauser PM, Bernard T, Greub G, Jaton K, Pagni M, Hafen GM (2014) Microbiota present in cystic fibrosis lungs as revealed by whole genome sequencing. PLoS One 9(3):e90934.

18. Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, Abbott K, Edwards R, Evangelista J, Conrad D, Rohwer F (2013) Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities. J Cyst Fibros 12(2):154-64.

19. Lim YW, Evangelista JS 3rd, Schmieder R, Bailey B, Haynes M, Furlan M, Maughan H, Edwards R, Rohwer F, Conrad D (2014) Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. J Clin Microbiol 52(2):425-37.

20. Streit WR, Schmitz RA (2004) Metagenomics--the key to the uncultured microbes. Curr Opin Microbiol 7(5):492-8. Review.

21. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5(1):16-8.

22. Chistoserdova L (2010) Recent progress and new challenges in metagenomics for biotechnology. Biotechnol Lett 32(10):1351-9.

23. Wooley JC, Ye Y (2009) Metagenomics: Facts and Artifacts, and Computational Challenges*. J Comput Sci Technol 25(1):71-81.

24. Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2(1):3.

25. Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6(2):e1000667.

26. J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5(10):R245–R249.

27. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72(4):557-78.

28. Davenport CF, Tümmler B (2013) Advances in computational analysis of metagenome sequences. Environ Microbiol 15(1):1-5.

29. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J; MetaHIT Consortium, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59-65.

30. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics, 9:386.

31. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40(Database issue):D115-22.

32. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40(Database issue):D571-9.

33. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36(Database issue):D13-21.

34. Woese CR, Stackebrandt E, Macke TJ, Fox GE (1985) A phylogenetic definition of the major eubacterial taxa. Syst Appl Microbiol 6:143-51.

35. Woese CR (1987) Bacterial evolution. Microbiol Rev 51(2):221-71. Review.

36. Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 104(27):11436-40.

37. Coburn B, Wang PW, Diaz Caballero J, Clark ST, Brahma V, Donaldson S, Zhang Y, Surendra A, Gong Y, Elizabeth Tullis D, Yau YC, Waters VJ, Hwang DM, Guttman DS (2015) Lung microbiota across age and disease stage in cystic fibrosis. Sci Rep 5:10241.

38. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG (2014) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol 16(9):2659-71.

39. Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J 3(12):1365-73.

40. Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, Eisen JA, Pollard KS (2011) PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Comput Biol 7(1):e1001061.

41. Schloss PD (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 6(7):e1000844.

42. Jumpstart Consortium Human Microbiome Project Data Generation Working Group (2012) Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One 7(6):e39315.

43. Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10(5):490-8.

44. Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21(24):4320-1.

45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403-10.

46. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755-63. Review.

47. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94(3):441-8.

48. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641-58.

49. Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19(R2):R227-40.

50. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299(5607):682-6.

51. Schneider GF, Dekker C (2012) DNA sequencing with nanopores. Nat Biotechnol 30(4):326-8.

52. Sibley CD, Grinwis ME, Field TR, Eshaghurshan CS, Faria MM, Dowd SE, Parkins MD, Rabin HR, Surette MG (2011) Culture enriched molecular profiling of the cystic fibrosis airway microbiome. PLoS One 6(7):e22702.

53. Bittar F, Rolain JM (2010) Detection and accurate identification of new or emerging bacteria in cystic fibrosis patients. Clin Microbiol Infect 16(7):809-20.

54. Carmody LA, Zhao J, Schloss PD, Petrosino JF, Murray S, Young VB, Li JZ, LiPuma JJ (2013) Changes in cystic fibrosis airway microbiota at pulmonary exacerbation. Ann Am Thorac Soc 10(3):179-87.

55. Develioglu ON, Ipek HD, Bahar H, Can G, Kulekci M, Aygun G (2014) Bacteriological evaluation of tonsillar microbial flora according to age and tonsillar size in recurrent tonsillitis. Eur Arch Otorhinolaryngol 271(6):1661-5.

56. Zemanick ET, Harris JK, Wagner BD, Robertson CE, Sagel SD, Stevens MJ, Accurso FJ, Laguna TA (2013) Inflammation and airway microbiota during cystic fibrosis pulmonary exacerbations. PLoS One 8(4):e62917.

57. Folkesson A, Jelsbak L, Yang L, Johansen HK, Ciofu O, Høiby N, Molin S (2012) Adaptation of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective. Nat Rev Microbiol 10(12):841-51.

58. Taylor CJ, McGaw J, Howden R, Duerden BI, Baxter PS (1990) Bacterial reservoirs in cystic fibrosis. Arch Dis Child 65(2):175-7.

59. Cramer N, Wiehlmann L, Ciofu O, Tamm S, Høiby N, Tümmler B (2012) Molecular epidemiology of chronic Pseudomonas aeruginosa airway infections in cystic fibrosis. PLoS One 7(11):e50731.

60. Cantón R, Fernández Olmos A, de la Pedrosa EG, del Campo R, Antonia Meseguer M (2011 [Chronic bronchial infection: the problem of Pseudomonas aeruginosa]. Arch Bronconeumol 47 Suppl 6:8-13.

61. Proctor RA, von Eiff C, Kahl BC, Becker K, McNamara P, Herrmann M, Peters G (2006) Small colony variants: a pathogenic form of bacteria that facilitates persistent and recurrent infections. Nat Rev Microbiol 4(4):295-305. Review.

62. Goerke C, Wolz C (2010) Adaptation of Staphylococcus aureus to the cystic fibrosis lung. Int J Med Microbiol 300(8):520-5.

63. Cakır Aktas N, Erturan Z, Karatuna O, Karahasan Yagci A (2013) Panton-Valentine leukocidin and biofilm production of Staphylococcus aureus isolated from respiratory tract. J Infect Dev Ctries 7(11):888-91.

64. Jones AM, Dodd ME, Webb AK (2001) Burkholderia cepacia: current clinical issues, environmental controversies and ethical dilemmas. Eur Respir J 17(2):295-301. Review.

65. Waters V, Yau Y, Prasad S, Lu A, Atenafu E, Crandall I, Tom S, Tullis E, Ratjen F (2011) Stenotrophomonas maltophilia in cystic fibrosis: serologic response and effect on lung disease. Am J Respir Crit Care Med 183(5):635-40.

66. Starner TD, Zhang N, Kim G, Apicella MA, McCray PB Jr (2006) Haemophilus influenzae forms biofilms on airway epithelia: implications in cystic fibrosis. Am J Respir Crit Care Med 174(2):213-20.

67. Ochman H, Moran NA (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science 292(5519):1096-9. Review.

68. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784):299-304. Review.

69. Doolittle RF, Feng DF, Anderson KL, Alberro MR (1990) A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J Mol Evol 31(5):383-8.

70. Didelot X, Maiden MC (2010) Impact of recombination on bacterial evolution. Trends Microbiol 18(7):315-22.

71. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98(1):182-7.

72. Hao W (2013) Extensive genomic variation within clonal bacterial groups resulted from homologous recombination. Mob Genet Elements 3(1):e23463.

73. Spratt BG (2004) Exploring the concept of clonality in bacteria. Methods Mol Biol 266:323-52. Review.

74. Perron GG, Lee AE, Wang Y, Huang WE, Barraclough TG (2012) Bacterial recombination promotes the evolution of multi-drug-resistance in functionally diverse populations. Proc Biol Sci 279(1733):1477-84.

75. Normark BH, Normark S (2002) Evolution and spread of antibiotic resistance. J Intern Med 252(2):91-106. Review.

76. Pollack M (1984) The virulence of Pseudomonas aeruginosa. Rev Infect Dis 6 Suppl 3:S617-26. Review.

77. Tümmler B, Wiehlmann L, Klockgether J, Cramer N (2014) Advances in understanding Pseudomonas. F1000Prime Rep 6:9.

78. Matz C, Bergfeld T, Rice SA, Kjelleberg S (2004) Microcolonies, quorum sensing and cytotoxicity determine the survival of Pseudomonas aeruginosa biofilms exposed to protozoan grazing. Environ Microbiol 6(3):218-26.

79. Mah TF, Pitts B, Pellock B, Walker GC, Stewart PS, O'Toole GA (2003) A genetic basis for Pseudomonas aeruginosa biofilm antibiotic resistance. Nature 426(6964):306-10.

80. Landry RM, An D, Hupp JT, Singh PK, Parsek MR (2006) Mucin-Pseudomonas aeruginosa interactions promote biofilm formation and antibiotic resistance. Mol Microbiol 59(1):142-51.

81. Hilker R, Munder A, Klockgether J, Losada PM, Chouvarine P, Cramer N, Davenport CF, Dethlefsen S, Fischer S, Peng H, Schönfelder T, Türk O, Wiehlmann L, Wölbeling F, Gulbins E, Goesmann A, Tümmler B (2015) Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment. Environ Microbiol 17(1):29-46.

82. He J, Baldini RL, Déziel E, Saucier M, Zhang Q, Liberati NT, Lee D, Urbach J, Goodman HM, Rahme LG (2004) The broad host range pathogen Pseudomonas aeruginosa strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. Proc Natl Acad Sci U S A 101(8):2530-5.

83. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Köhler T, van Delden C, Weinel C, Slickers P, Tümmler B (2007) Population structure of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 104(19):8101-6.

84. Wiehlmann L, Munder A, Adams T, Juhas M, Kolmar H, Salunkhe P, Tümmler B (2007) Functional genomics of Pseudomonas aeruginosa to identify habitat-specific determinants of pathogenicity. Int J Med Microbiol 297(7-8):615-23.

85. Dinesh SD, Grundmann H, Pitt TL, Römling U (2003) European-wide distribution of Pseudomonas aeruginosa clone C. Clin Microbiol Infect 9(12):1228-33.

86. Römling U, Kader A, Sriramulu DD, Simm R, Kronvall G (2005) Worldwide distribution of Pseudomonas aeruginosa clone C strains in the aquatic environment and cystic fibrosis patients. Environ Microbiol 7(7):1029-38.

87. Cramer N, Klockgether J, Wrasman K, Schmidt M, Davenport CF, Tümmler B (2011) Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs. Environ Microbiol 13(7):1690-704.

88. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing (2008) Nucleic Acids Res 36(16):e105.

89. Tamames J, Moya A (2008) Estimating the extent of horizontal gene transfer in metagenomic sequences. BMC Genomics 9:136.

90. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523(7559):208-11.

91. Oh J, Byrd AL, Deming C, Conlan S; NISC Comparative Sequencing Program, Kong HH, Segre JA (2014) Biogeography and individuality shape function in the human skin metagenome. Nature 514(7520):59-64.

92. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto JM, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J (1012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490(7418):55-60.

93. Pace NR (2009) Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev 73:565–76.

94. Grice EA, Kong HH, Renaud G, et al. (2008) A diversity profile of the human skin microbiota. Genome Res 18:1043–50.

95. Delgado S, Suárez A, Mayo B (2006) Identification of dominant bacteria in feces and colonic mucosa from healthy Spanish adults by culturing and by 16S rDNA sequence analysis. Dig Dis Sci 51:744–51.

96. Hold GL, Pryde SE, Russell VJ, et al. (2002) Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis. FEMS Microbiol 39:33–9.

97. Barlow GM, Yu A, Mathur R (2015) Role of the Gut Microbiome in Obesity and Diabetes Mellitus. Nutr Clin Pract 9. Review.

98. Hartstra AV, Bouter KE, Bäckhed F, Nieuwdorp M (2015) Insights into the role of the microbiome in obesity and type 2 diabetes. Diabetes Care 38(1):159-65. Review.

99. Coburn B, Wang PW, Diaz Caballero J, Clark ST, Brahma V, Donaldson S, Zhang Y, Surendra A,

Gong Y, Elizabeth Tullis D, Yau YC, Waters VJ, Hwang DM, Guttman DS (2015) Lung microbiota across age and disease stage in cystic fibrosis. Sci Rep 5:10241.

100. Rabin HR, Surette MG (2012) The cystic fibrosis airway microbiome. Curr Opin Pulm Med 18(6):622-7. Review.

101. Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, Bruce KD (2004) Characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16s ribosomal DNA terminal restriction fragment length polymorphism profiling. J Clin Microbiol 42(11):5176-83.

102. Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P (2013) Metagenomics for pathogen detection in public health. Genome Med 5(9):81. Review.

103. Pallen MJ (2014) Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. Parasitology 141(14):1856-62.

104. Sears CL, Garrett WS (2014) Microbes, microbiota, and colon cancer. Cell Host Microbe 15(3):317-28. Review.

105. Irrazábal T, Belcheva A, Girardin SE, Martin A, Philpott DJ (2014) The multifaceted role of the intestinal microbiota in colon cancer. Mol Cell 54(2):309-20. Review.

106. Kostic AD, Xavier RJ, Gevers D (2014) The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology 146(6):1489-99. Review.

107. Liang J, Sha SM, Wu KC (2014) Role of the intestinal microbiota and fecal transplantation in inflammatory bowel diseases. J Dig Dis 15(12):641-6. Review.

108. Gilbert JA, Meyer F, Bailey MJ (2011) The future of microbial metagenomics (or is ignorance bliss?). ISME J 5(5):777-9.

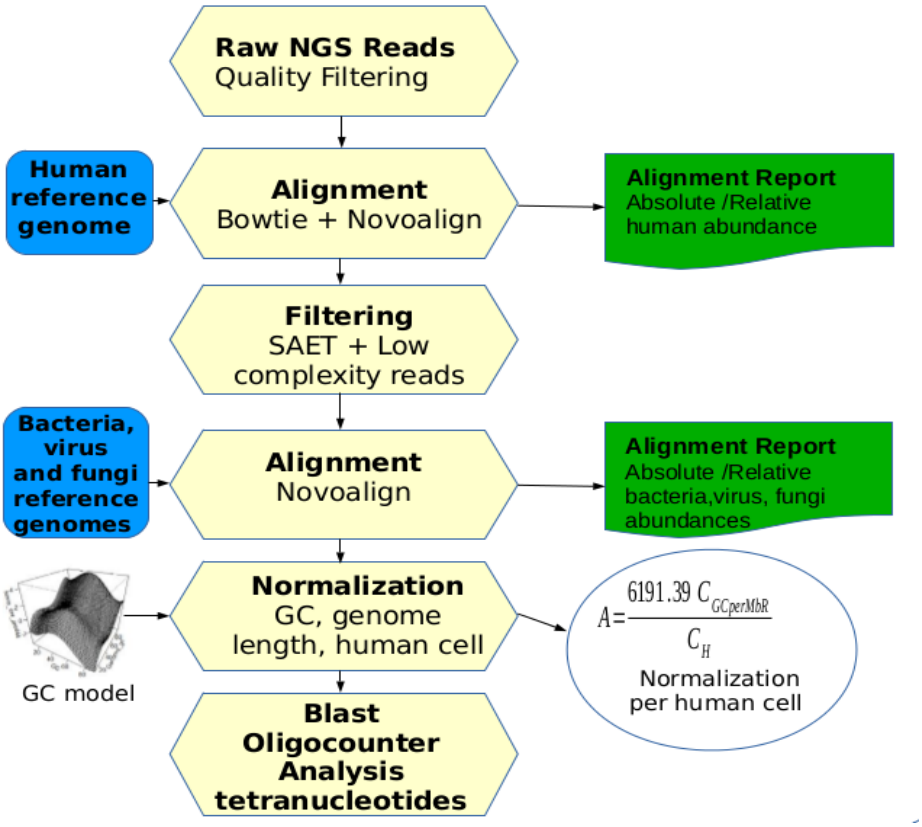
109. Pible O, Armengaud J (2015) Improving the quality of genome, protein sequence, and taxonomy databases: A prerequisite for microbiome meta-omics 2.0. Proteomics 15(20):3418-23.

Chapter 10

Appendix

Appendix1 - Metagenomic analysis pipeline

Here the metagenomic pipeline is described a bit more in detail:



1. **Quality Filtering:** sequences are trimmed and filtered out based on the quality parameters. The command line used is:

   python solid-trimmer.py -c input -q output -p outfile.trimmed --max-ns 3 --moving-average 7:18 --min-read-length 45

2. **Alignment against human reference genome:** quality filtered sequences are aligned against the human reference genome, first with the fast aligner Bowtie2:

```
bowtie2 --fast -x  human.reference -U  input.trimmed.fastq -S output.sam
```

and secondly, the remaining sequences are aligned again against the human reference genome. This second alignment is performed with the specific aligner for short reads called Novoalign:

```
novoalignCS -d human.reference.cix -f input.csfasta –F CSFASTAnQV  -r Random  -H -c 24 -o SAM
```

3. **Filtering of low complexity reads:** the extraction of low complexity reads is achieved to all non-human sequences with the tool prinseq-lite as following:

```
./prinseq-lite-0.20.4/prinseq-lite.pl -fastq input -lc_method dust -lc_threshold 5 -out_good null -out_bad outfile.bad
```

4. **Bacteria alignments:** to perform the taxonomic bacterial classification two different steps of alignments are executed with the software Novoalign.

   1st Step: the option –r None is selected in order to identify unique sequences, i.e. sequences that match uniquely against a specific region. The command used is:

   ```
   novoalignCS -d bacteria.genomes.cix –f input.filtered.csfasta -F CSFASTAnQV -r None  -H -c 24 -o SAM
   ```

   2nd Step: the option –r All is used to identify all possible hits against bacterial genomes.

   ```
   novoalignCS -d bacteria.genomes.cix –f input.filtered.csfasta -F CSFASTAnQV -r All  -H -c 24 -o SAM
   ```

5. **Virus and Fungi alignments:** all remaining sequences are aligned against DNA viruses and Fungi reference genomes with the aligner Novoalign.

```
novoalignCS -d virus_fungi_reference.cix –f  input.no_bacteria.csfasta -F CSFASTAnQV -r None -H -c 24 -o SAM
```

6. **Normalization:** normalization is performed based on the model explained in Chapter 6. All sequences aligned against bacteria, DNA viruses, fungi and molds are normalized by GC content and genome length before relative abundances are reported.
   As well, a step of normalization by human cell is performed as is described in Chapter 6.

7. **MetaPhlAn2:** this program is used to calculate the heatmaps. The following commands were used:

```
merge_metaphlan_tables.py *.txt > merged_abundance_table.txt

metaphlan_hclust_heatmap.py -c bbcry --top 10 --minv 0.01 -s log –in
merged_abundance_table.txt --out  abundance_heatmap_top10.png
```

For supplementary R scripts developed during this thesis, please refer to the DVD attached to the thesis.

Appendix 2 - Haplotypes reconstruction pipeline

A new algorithm was developed to perform the analysis of recombination based on haplotypes reconstruction. The script and an example file are provided with the DVD attached to the thesis. The script is written in perl and performs pair-comparisons of SNPs detected in each strain of the analysis. The input provides the information of SNPs positions for each strain, an example of the input file format is:

Strain	SNP_position	nt	nt_variant
0812	21656	T	A
0812	34371	T	C
0812	34689	A	G
0812	39372	G	A
1BAE	34371	T	C
1BAE	49639	T	C
1BAE	57196	A	G
1BAE	57217	G	A
239A	34371	T	C
239A	64139	T	C
239A	74468	C	G

Different output files will be generated:

- all_files_SNPs.txt: contains all different SNPs for each strain.
- combinations.txt: provides the information of all different strain comparisons.
- haplotypes_distances_hp.txt: this file has the information regarding all haplotypes found in all strain comparisons. Each row contains the information of a haplotype. Each column contains:
    - 1st column: number of haplotypes.
    - 2nd column: number of consecutive SNPs that two strains share in that haplotype.
    - 3rd column: all SNPs were sorted by increasing genome position. Rank numbers were assigned where 0 represents the first SNP position found. The start SNP rank number of the haplotype is found in this column.
    - 4th column: End SNP rank number of the haplotype.
    - 5th column: represents the start position of the haplotype in the reference genome.
    - 6th column: end position of the haplotype in the reference genome.
    - 7th column: genome physical length of the haplotype.
    - 8th column: strains to which the haplotype belongs.

Number_ haplotype	Consecutive _SNPs	SNP_ start	SNP_ end	Hb_genome _st	Hb_genome_ end	Genome _length	Comparison
1	4	0	3	16654	34689	18036	E429_EC21
2	4	6	9	29639	61809	12171	E429_EC21
3	5	15	19	94590	115976	21387	E429_EC21
4	2	21	22	132188	133223	1036	E429_EC21
5	2	25	26	143876	155532	11657	E429_EC21
6	3	28	30	159798	164956	5159	E429_EC21
7	19	35	53	190098	285599	95502	E429_EC21

- TOTAL_MATRIX.txt: The first line of the file represents the SNPs positions in the reference genome. The second line shows the rank number of the SNPs detected. Following lines contain the information of all consecutive SNPs at each position for each strain comparison. Example of the file:

Genome position ->	35	60	432	765	854	996
SNP position ->	0	1	2	3	4	5
Comparison1 ->	3	2	1	0	2	1
Comparison2 ->	5	4	3	2	1	0

Abbreviations

16S rRNA	16S ribosomal RNA
bp	nucleotide base pair
Bcc	*Burkholderia cepacia* complex
CF	cystic fibrosis
ddNTP	dideoxynucleotides
DNA	deoxyribonucleic acid
emPCR	emulsion PCR
GC	G+C content, the percentage of summed guanine and cytosine nucleotides in a sequence
MRSA	Methicillin-resistant Staphylococcus aureus
NGS	next generation sequencing
NTHi	Unencapsulated *Haemophilus influenzae* strains
PI	pancreatic insufficient
PS	pancreatic sufficient
PCR	polymerase chain reaction
Q	quality sequencing value
rRNA	ribosomal RNA
SNP	single nucleotide acid
WGS	whole-genome shotgun sequencing

Contributions

A list of colleagues and their contributions to this project are outlined below.

Angela Schulz and Silke Hedtfeld performed the extraction of specimens and DNA preparation.
Marie Dorda, Samira Mielke and Lutz Wiehlmann performed the library preparation and sequencing of samples
Philippe Chouvarine provided bioinformatic support.

Curriculum vitae

Patricia Morán Losada

Vahrenwalder Str.34

30519 Hannover

08/08/1982
Carracedelo (Spain)

Tlf. +49 176 87782083

patriciamoranl@hotmail.es

Education

Qualification: PhD – Program Infection biology (actual)
Place of Study: Clinic for Paediatric Pneumology and Neonatology, Hannover Medical School. Working group of Prof. Dr. rer.nat. Burkhard Tümmler.
Duration: Since October 2012.

Qualification: Masters in Bioinformatics & Computational Biology.
Place of Study: "Universidad Complutense" of Madrid.
Duration: October 2010-July 2011.

Qualification: Bachelor of Biology
Place of Study: University of Leon (Last year completed in the University of Siena, Italy).
Duration: September 2000- September 2005.

Work experience

Company: **Lifesequencing (Valencia, Spain)**
Position: **Bioinformatician**
Duration: May 2011 – September 2012
Description: Analysis of next generation sequencing data produced by 454, Illumina and Ion Torrent technologies.
Management of environmental research projects, as well as human genetic studies.

Company: **F. Iniciatives –R&D (Madrid, Spain)**
Position: **Project Management**
Duration: January 2011 – April 2011
Description: Development of R&D. The projects carried out were related to the environmental sector.

Company: **Wellness Telecom S.L (Valladolid, Spain)**
Position: **Project Management – R&D**
Duration: January 2010 - September 2010
Description: Conducting R&D projects.
The projects ranged from national to European and international.

Company: **ADEuropa (Valladolid, Spain)**
Position: **Internship**
Duration: September – December 2009
Description: R & D project management.

Company: **Accenture -Corited (Madrid, Spain)**
Position: **Programmer**
Duration: April 2007 – September 2008
Description: Maintenance and programming new tools.

List of publications

- **Losada PM**, Chouvarine P and Tümmler B. Bacterial recombination analysis based on haplotype construction. [Submitted].
- Hilker R, Munder A, Klockgether J, **Losada PM**, Chouvarine P, Cramer N, Davenport CF, Dethlefsen S, Fischer S, Peng H, Schönfelder T, Türk O, Wiehlmann L, Wölbeling F, Gulbins E, Goesmann A, Tümmler B. Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment. [Environ Microbiol, 2015].
- Fischer S, Cramer N, **Losada PM**, Chouvarine P, Dethlefsen S, Davenport C, Dorda M, Goesmann A, Hilker R, Mielke S, Schönfelder T, Suerbaum S, Türk O, Woltemate S, Wiehlmann L, Klockgether J and Tümmler B. Intraclonal genome diversity of the major *Pseudomonas aeruginosa* clones C and PA14. [Environmental Microbiology: in revision].
- Chouvarine P, Wiehlmann L, **Losada PM** and Tümmler B. Filtration and normalization of sequencing read data in whole-metagenome shotgun samples. [Submitted].
- **Losada PM**, Chouvarine P, Dorda M, Hedtfeld S, Mielke S, Schulz A, Wiehlmann L, Tümmler B. The cystic fibrosis lower airways microbial metagenome. [Submitted].
- Saran S, Tran DD, Klebba-Färber S, **Moran-Losada P**, Wiehlmann L, Koch A, Chopra H, Pabst O, Hoffmann A, Klopfleisch R, Tamura T (2013) THOC5, a member of the mRNA export complex, contributes to processing of a subset of wingless/integrated (Wnt) target mRNAs and integrity of the gut epithelial barrier. BMC Cell Biol 14:51.

Declaration

Herewith, I confirm that I have written the present PhD thesis myself and independently, and that I have not submitted it at any other university worldwide.

Hannover, 17th November 2015

_____