

Genomic and metagenomic application of microbial genome signatures



**A thesis accepted for the degree of
Doctor of Philosophy (PhD)
in the subject of Bioinformatics**

by

Colin Davenport, MSc Bioinformatics, BSc Ecology

September 2009

**Hanover Medical School
International Research Training Group "Pseudomonas"
in Hanover Biomedical Research School (HBRS)
*Department of Paediatrics***

Acknowledged by the PhD committee and head of Hanover Medical School

President: Prof. Dr. med. Dieter Bitter-Suermann

Supervisor: Prof. Dr. Dr. Burkhard Tümmler

Cosupervisor: Dr. Oleg Reva

External expert: Prof. David W. Ussery

External expert: Prof. Peter Golyshin

Day of public defense: 24.11.2009

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.

Hanover, (September 2009)

I believe that we have been doing this not primarily to achieve riches or even honour, but rather because we were interested in the work, enjoyed doing it and felt very strongly that it was worthwhile.

Frederick Sanger

Hell is sitting on a hot stone and being forced to read your own scientific publications.

Erik Ursin, fish biologist

Acknowledgements

Special thanks are due to:

my supervisor Burkhard Tümmler, without whom none of this work would ever have been started, let alone completed.

my co-supervisor Oleg Reva for his considerable interest and input during two visits to South Africa and at several conferences.

Prof. David Ussery for reviewing this work and examining my presentation.

Prof. Peter Golyshin, also for reviewing this thesis and acting as an examiner.

the Clinical Research Group at Hanover Medical School, in particular Jens Klockgether, Frauke Stanke and Lutz Wiehlmann, for their interest and support for this work, despite not even pretending to be bioinformaticians.

the bioinformaticians at the Computational Biology research unit in Pretoria, and additionally J. Ringe and C. Schmidt at Hanover Medical School, all for their invaluable assistance in technical matters.

My family, for putting up with me while I was completing and writing this work.

Summary

This thesis focuses on the application of patterns in DNA termed oligonucleotide signatures in genomic and metagenomic research. Due to the recent rapid advances in sequencing technologies costs have decreased to the point that microbial sequencing is a routine procedure. To gain information from this data comparative genomics using various computational tools is required. Short to medium sized oligomers (4-14 bp) were demonstrated to be useful for analysis of core and atypical genomic regions in the genus *Pseudomonas*. In other words, oligonucleotide usage varied between the core genome and accessory elements such as genomic islands which have been putatively acquired by horizontal gene transfer. Distributions of these oligomer copy numbers or parameters were best assessed via two computational visualisation tools made available to the community as part of the OligoCounter package. Furthermore, globally distributed mid-length oligomers of sufficient specificity to be of use for assignment of taxonomic origin of DNA fragments were localised, and were depleted in putatively horizontally transferred regions. Oligonucleotide usage from the web-based SeqWord tool was also shown to be of similar utility in investigation of the phylum Chlorobi (Green Sulfur Bacteria). Additional whole proteome bioinformatic comparisons supported, consistent with previous work, the taxonomic reassignment of multiple species within the Chlorobi. Interestingly, taxonomy in this group has been historically based on phenotypic characteristics, whereas molecular assessment has caused sweeping changes. Taxonomy was also key to the discovery of a class of abundant near complementary oligomers overrepresented throughout the prokaryotes. These distinctive oligomers, composed of a polypurine tract with a central adenine, were not biased towards a reading frame when in coding regions and displayed extremal structural parameters. A role in triple helix formation was proposed based on these characteristics. Finally, modern short read sequencing data from several *P. aeruginosa* Illumina resequencing projects were evaluated. The large number of short reads led to accurate resolution of SNPs and insights into microevolution in the laboratory and human lung habitats. Short reads were also found to contain sufficient phylogenetic signal for taxonomic attribution of reads to species level in metagenomics. Numerous published metagenomes were rapidly reanalysed by the Metatite system which sometimes led to new findings, due to Metatite's specificity and the recent increase in sequenced organisms. The work was wholly computational in nature and was performed at Hanover Medical School and the Bioinformatics facility at the University of Pretoria in South Africa.

Keywords: metagenomics, short reads, oligonucleotide usage, comparative genomics

Zusammenfassung

Die vorliegende Arbeit beschreibt die Anwendung von DNA Mustern, sogenannten Oligonukleotidsignaturen, in der genomischen und metagenomischen Forschung. Durch technische Neuerungen im Bereich der Sequenzieretechnologien wurden die Kosten für eine Sequenzierung soweit reduziert, dass die Sequenzierung von Bakterien Routine wurde. Um aus diesen Daten Nutzen zu ziehen, bedarf es der computergestützten, vergleichenden Genomanalyse. Kurze bis mittellange Oligomere (4-14bp) haben ihren Nutzen in der Analyse des Kerngenoms und atypischer Genombereiche in der Gattung *Pseudomonas* bewiesen. Die Oligonukleotidnutzung variierte zwischen dem Kerngenom und akzessorischen Elementen, zum Beispiel Geninseln, welche durch horizontalen Gentransfer erworben wurden. Die Verteilung der Kopienzahlen der Oligomere wurden am effektivsten beschrieben mit Hilfe zweier Visualisierungshilfsmittel, die der Forschergemeinschaft als Teil des OligoCounter Softwarepakets zugänglich gemacht wurden. Global verteilte mittellange Oligomere, die eine ausreichende Spezifität besitzen um den taxonomischen Ursprung des DNA Fragments zu bestimmen, wurden lokalisiert, fehlten aber in Regionen welche potenziell durch horizontalen Gentransfer akquiriert wurden. Oligonukleotidnutzung, ermittelt mit Hilfe des web-basierten SeqWord Tools, zeigte außerdem Anwendbarkeit für das Phylum Chlorobi (Grüne Schwefelbakterien). Zusätzlich wurden mit Hilfe bioinformatischer Vergleiche des gesamten Proteoms sowie mit Zuhilfenahme vorheriger Arbeiten mehrere Spezies innerhalb der Gattung Chlorobi taxonomisch neu zugeordnet. Taxonomie wurde in dieser Gattung historisch bedingt auf der Basis von phänotypischen Merkmalen klassifiziert, eine molekulare Untersuchung der einzelnen Spezies führte zu einigen Änderungen. Die Taxonomie war auch der Schlüssel für die Entdeckung einer Klasse von vielen, fast komplementären Oligomeren, welche in den Prokaryoten überrepräsentiert sind. Diese speziellen Oligomere, welche aus einer Polypurinfolge mit einem zentralen Adenin bestehen, waren nicht auf ein Leseraster beschränkt und zeigten extreme strukturelle Eigenschaften. Eine Rolle in der Bildung von Triplehelixen wurde basierend auf diesen Eigenschaften prognostiziert. Abschließend wurden Sequenzierungsdaten von mehreren *P. aeruginosa* Resequenzierungsprojekten analysiert. Die große Anzahl kurzer Reads führte zur Auflösung von SNPs und ermöglichte Einblicke in die Mikroevolution im Labor und der menschlichen Lunge. Kurze Reads zeigten außerdem genügend phylogenetische Marker, um eine metagenomische Spezieseinteilung zu gewährleisten. Eine Vielzahl publizierter Metagenome wurde in kurzer Zeit mittels der Metatier Software reanalysiert, was in einigen Fällen zu neuen Erkenntnissen führte, was auf die Spezifität von Metatier und die Vielzahl neusequenzierter Mikroorganismen zurückzuführen ist.

Schlüsselwörter: vergleichende Genomanalyse, Genomsignaturen, kurze Reads, Metagenomik

Table of contents

Introduction

1.	History	I-1
2.	Recent progress in DNA sequencing technologies	I-1
3.	Comparative genomics	I-2
4.	Genomic islands and anomalous genomic elements	I-3
5.	Phenotypic relevance of the accessory genome	I-4
6.	Oligonucleotide analysis	I-4
7.	Metagenomics	I-6
8.	Projects in Metagenomics: the State of the Art	I-6
9.	Challenges in metagenomics	I-8
10.	References	I-9

Objectives	O-1
-------------------	-----

Results

1.	Comparative genomics of the genus <i>Pseudomonas</i> and phylum Chlorobi	
1.1	Background	1-1
1.1.1	Comparative genomics	1-1
1.1.2	Oligonucleotide usage	1-2
1.2	About the manuscripts	1-3
	Davenport et al. 2009, <i>Environ Microbiol</i> 11 , 1092-1104.	
	Davenport et al., submitted to <i>Photosynthesis Research</i> .	
1.3	References	1-5
2.	The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage	
2.1	Background	2-1
2.2	About the manuscript	2-1
	Ganesan et al. 2009, <i>BMC Bioinformatics</i> 9 , 333.	
2.3	References	2-2

3.	Global distributions of oligomers across the prokaryotic kingdom	
3.1	Background	3-1
3.1.1	Measuring overrepresentation	3-1
3.2	About the manuscripts	3-2
	Davenport & Tümmeler, in revision at <i>Plos One</i> .	
	Emmett et al., in revision at <i>Environmental Microbiology</i> .	
3.3	References	3-3
4.	Genome Diversity of <i>Pseudomonas aeruginosa</i> PAO1 laboratory strains	
4.1	Background	4-1
4.1.1	Genome assembly with short reads	4-1
4.2	About the manuscript	4-2
	Klockgether et al., in revision at the <i>Journal of Bacteriology</i> .	
4.3	References	4-3
5.	Metatie - a fast, accurate and adaptable metagenomic classifier for ultra-short reads	
5.1	Background	5-1
5.1.1	Short reads in metagenomics	5-2
5.2	About the manuscript	5-2
	Davenport et al. 2009, submitted to <i>Genome Research</i> .	
5.3	References	5-4
Conclusions and outlook		
1	Oligonucleotide signature analysis	C-1
2	Comparative genome analysis	C-2
3	<i>In silico</i> metagenomic analysis	C-4
4	Short read assembly	C-6
5	References	C-7

Abbreviations

Appendix

Introduction

1. History

Over the course of the last half century, much of the progress in microbiology has gone hand in hand with advances in molecular biology. Microorganisms, while known to be relevant pathogens and commensals since the days of Koch and before, had proved difficult to detect, isolate, culture, and monitor. These problems arose due to prokaryote size, habitats, life histories, nutrient requirements, and rapid population dynamics, among others. For example, completely unrelated bacteria may be very similar in phenotype which can greatly confuse identification. Methods of directly assessing and investigating bacteria on a genetic level were necessary to properly evaluate and build on phenotypic observations. Following identification of the role of deoxyribose nucleic acid as the main molecule of information storage (Avery et al. 1944) and elucidation of its molecular structure (Watson and Crick, 1953), research on microbes quickly progressed. One example is the use of molecular biological techniques to unravel the regulation of the *lac* operon (Jacob and Monod 1961) in the 1960s. Microbial genomes also proved amenable, due to their small size and comparative lack of repeats, to initial progress with the nucleotide sequencing technologies in the 1970s. These methods were introduced by Sanger and coworkers (Sanger and Coulson 1975) and others (Gilbert and Maxam 1973). The 3569 bp RNA bacteriophage MS2 was the first organism to be completely sequenced (Fiers et al. 1976), with the first DNA sequence completed by Sanger and coworkers soon after (Sanger et al. 1977). Thereafter followed in the next two decades a range of incremental increases in sequencing capacity. These achievements culminated in the completion of the first genome of a free living organism, *Haemophilus influenzae*, in 1995 (Fleischmann et al. 1995). Other prokaryotes such as *Escherichia coli* soon followed. Among the first eukaryotes to be sequenced were bakers yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster* and the considerable international efforts which led to the *H. sapiens* genomic sequence (Lander et al. 2001).

2. Recent progress in DNA sequencing technologies

Like many disciplines within the natural sciences, waves of discovery are closely associated with and facilitated by novel technical advances. The world of molecular biology is currently experiencing a pulse of discovery aided and abetted by the so-called next generation sequencing

technologies. Up until about 2005, sequencing had been almost exclusively reliant on the established Sanger sequencing methodology (Margulies et al. 2005). Even large scale sequencing projects such as the Human Genome project had been carried out using these relatively slow and expensive technology (Venter et al. 2001). Typically, large numbers of sequencers were continuously operated in parallel by teams of technical staff in order to generate sufficient sequence. Such expense and effort placed whole genome sequencing in the realm of international cooperations and large facilities, well beyond the reach of typical research laboratories. That all changed with the introduction of the first of the so called next generation sequencers in 2005 (Margulies et al. 2005). Suddenly several orders of magnitude more sequence data could be produced in a fraction of the time and cost. This device was only the first of multiple devices, with further massively parallel sequencers arriving from Illumina, Helicos, and Applied Biosystems, among others. These devices have enabled single laboratories access to genomic information at rapidly decreasing prices, and this trend looks set to continue (Hugenholtz and Tyson 2008). For example, a single molecule sequencing approach was recently utilised by three workers to sequence an entire human genome at 28 fold coverage for approximately 50000 US dollars in four weeks (Pushkarev et al. 2009). Possible applications of these technologies are deep sequencing of environmental microbes without having to create clone libraries (Edwards and Rohwer 2005), in high throughput DNA methylation analysis for the purpose of assessing gene expression (Schaefer et al. 2009) or in discovery of residual variants of evolving cancer cell subclones at frequencies as low as 1 in 5000 (Campbell et al. 2008). Some applications even range as far as routine high throughput sequencing of genes or perhaps genomes in order to personalise medical treatments to the various patient genotypes (ten Bosch and Grody 2008). From a bioinformatic perspective the associated challenges include the technical issues involved in tracking, storing and analysing such large amounts of sequence data. At the same time the shorter read length of 25-400 bp has required development of a new generation of computational tools to be used effectively.

3. Comparative genomics

In the last decade the field of comparative genomics has come to prominence. This field focuses on deriving knowledge on diversity, pathogenicity and function from the genomes sequenced to date (Binnewies et al. 2006). Hundreds of completely sequenced genomes are available to compare data against and thousands more are in the process of being completed (Koonin and Wolf, 2008). Results have included advances on understanding of genome organisation, structure, and the widespread

nature of horizontal gene transfer (Ochman et al. 2005). For example, genomic regions including a so-called SCCmec island which influence the virulence and drug resistance of a methicillin resistant and a susceptible *Staphylococcus aureus* strain were identified by Holden and coworkers (2004) after sequencing. Comparative genomics has also implicated a decisive role for mobile accessory elements in prokaryote evolution (Hacker and Carniel 2001). While the current genome collections are undoubtedly highly biased towards human pathogens and easily culturable organisms, a range of concerted large scale efforts are underway to broaden knowledge of prokaryotes at a genomic level (the European Metahit program <http://www.sanger.ac.uk/pathogens/metahit/>; the Human Microbiome project, Turnbaugh 2007; the Genomic Encyclopedia for Bacteria and Archaea <http://www.jgi.doe.gov/programs/GEBA/why.html>; and the Moore Foundation Marine Microbial Genome Sequencing Project <http://www.moore.org/marine-micro.aspx>). These projects aim to sequence hundreds of microbes each from various habitats, and some are already well advanced.

4. Genomic islands and anomalous genomic elements

A bacterial genome can be divided into core and accessory regions (Juhas et al. 2009). The core genome is the "typical" part with GC content (van Passel et al. 2005) and oligomer usage typically fluctuating around a relatively consistent average value or pattern (Karlin et al. 1997, Reva and Tümmler 2004). These phenomena may arise due to constraints in replication and repair mechanisms, DNA structure and codon usage, among others (Pride et al. 2003). A range of genetic elements can insert into the core genome, thereby disrupting these nucleotide patterns. These include Insertion (IS) elements, transposons, bacteriophages, or genomic islands. Accessory elements can thus be distinguished by anomalous nucleotide content (see below), repeat sequences, proximity to tRNA genes, or by comparative genomics with syntenic regions in otherwise distantly related bacteria (Dobrindt et al. 2004).

Accessory regions also frequently encode mobility genes such as transposases and integrases (Dobrindt et al. 2004, Cortez et al. 2009). These mobility associated genes are crucial in allowing the element to excise and reintegrate into other chromosomes. Integration sites are typically located around tRNA genes, as these are an ever present and slowly evolving recognition target (Hacker and Carniel 2001). Larger mobile genetic elements themselves have a conserved backbone of replication associated genes, with smaller "cargo" regions carrying additional genes which are more likely to affect the phenotype and even fitness of the bacterium.

5. Phenotypic relevance of the accessory genome

The association between atypical genomic regions and pathogenicity was recognised early on, partially due to the historical tendency to sequence bacterial pathogens first. Pathogenicity factors such as toxins, adhesins or drug efflux transporters are commonly found in close association with mobile elements (Dobrindt et al. 2004). For example, the *Vibrio cholerae* SXT element is known to be associated with antibiotic resistance among a wide range of gram negative bacteria (Beaber et al. 2006). However, the link between pathogenicity and the presence of a genomic island cannot always be explained so easily. Non-pathogenic *E. coli* strains may contain identical or very similar islands to related highly pathogenic strains (Perna et al. 2001, Dobrindt et al. 2004). Fascinatingly, specific loss of genetic information can enhance bacterial virulence (Maurelli et al. 1998, Lee et al. 2006). Evolution of a genome, in particular by horizontal gene transfer, is directed by selection for an extra phenotypic ability to enable expansion into new niches or increase fitness in an existing ecological niche (Hacker and Carniel 2001, Dobrindt et al. 2004, Mathee et al. 2008). Genetic elements which confer a fitness advantage are likely to be stably integrated into the genome at some point, as they become common in the microbial population and mutations suppress action of mobility related genes (D'Argenio et al. 2007, Cramer et al., in preparation).

Some genomic islands not only excise at rapid rates but are highly promiscuous, integrating into genomes from completely different taxonomic lineages (Klockgether et al. 2007). Further, genes encoding a defined function are often clustered together, as several genes acting together may be required for a gain in fitness or phenotype (Cortez et al. 2009). For example, the genes of the *cob* operon in *Salmonella enterica* together encode vitamin B12 biosynthesis and were putatively acquired by *S. enterica* by horizontal gene transfer after divergence of *E. coli* (Vernikos et al. 2007). Another example is the existence of very similar clustered regularly interspaced short palindromic repeats (CRISPRs) in distantly related organisms (Godde and Bickerton 2006, Kunin et al. 2007). The green sulfur bacterium *Chlorobaculum parvum* NCIB 8327 contains CRISPRs which are not substantially different to those in Archaea. These structures have been linked to adaptive antiviral response mechanisms, and are thought to be ineffective if they are not acquired as a group with repeats, spacer sequences and associated proteins (Kunin et al. 2007).

6. Oligonucleotide analysis

Oligomer analysis is now a firmly established method for compositional investigation of nucleotide sequences (Karlin et al. 1997, Pride et al. 2003, Reva and Tümmler 2004, Worning et al. 2006,

Mrázek 2009, Willner et al. 2009). Put simply, oligomer analysis involves various methods for counting abundances and comparing patterns of di-, tri-, tetra- or higher order nucleotides from various sequences against each other. These patterns are likely the result of genome wide processes such as replication and repair or DNA structural biases (Karlin et al. 1997, Pride et al. 2003), or maintenance of regulatory machinery including transcription factor binding sites (Bush and Lahn 2006). Various workers have applied this compositional technique widely to investigate partial and complete genomes (Karlin et al. 1997), analyse phylogenetic relationships (Pride et al. 2003), discern core and accessory genomes (Reva and Tümmler 2004, van Passel et al. 2005, 2006, Ganesan et al. 2008), comparatively analyse large genomic collections (Bohlin et al. 2008) and even cluster and classify short pyrosequencing reads (Willner et al. 2009). For example, Bush and Lahn (2006) found similarities of 8 bp oligomers in regions immediately upstream of genes scale with evolutionary time. Furthermore, mid-length oligomers are automatically used behind the scenes by various models such as Interpolated Markov Models (Brady and Salzberg 2009). As oligonucleotide usage is a core section of many manuscripts and was extensively applied in the thesis, more detail is provided here.

GC content: While not actually an oligonucleotide based parameter, GC content serves to introduce the concept of a sliding window. The number of G and C base pairs are counted in a region of sequence (window) of, for example 10000 bp, results are listed as a percentage and the window slides on, typically by half the length of itself. This overlapping approach tends to reduce artifacts.

GC skew is an asymmetric nucleotide composition between leading and lagging strands, calculated as $(G-C)/(G+C)$ for a window of given length (Reva and Tümmler 2004, Marín and Xia 2008). Worning et al. (2006) found higher order oligomers up to 8mers to be optimal for precisely locating the origin of replication and terminus in bacterial chromosomes. This technique even works for chromosomes without extensively skewed preference of guanine to the leading strand (Worning et al. 2006).

Distance: Distance is a measure of divergence between two patterns of oligonucleotide frequencies, expressed as a percentage of the maximum possible distance. This is typically the distance between a pattern calculated for a local pattern versus that calculated for a whole genome.

Pattern skew: Pattern skew is very similar to Distance, it is namely the distance between oligomer patterns on the direct and reverse DNA strands of the same nucleotide sequence.

Oligonucleotide variance: Oligonucleotide variance is the numerical variance of

oligonucleotide word deviations of observed from expected counts. In general a higher value indicates a more highly biased and less random sequence. Low OUV scores in contrast indicate a sequence where the individual nucleotides in each tetranucleotide are less tightly dependent upon each other, which might occur where a large number of random mutations has taken place (Bohlin et al. 2008).

7. Metagenomics

Metagenomics is a new field of research which can be broadly defined as study of DNA gained directly from the environment (Hugenholtz and Tyson 2008). Metagenomics is of great interest as the currently well studied microbial organisms are those that are easily culturable, and these are the minority. The vast majority, perhaps as high as 99 % (Rappé and Giovannoni 2003), have largely evaded systematic investigation to date. Metagenomics provides a relatively unbiased (or at least non-hypothesis driven) view of community structure and metabolic potential of a community. Two main themes of research can be distinguished: screening of environmentally acquired libraries for enzymes displaying particular activity, i.e. functional metagenomics, or the application of DNA sequencing technology to environmental DNA sequences. While functional metagenomics has led to the discovery of a whole range of novel proteins, some of which are of great interest to industry due to particular thermostability or salt tolerance (Maróti et al. 2009), it is sequence based analyses of microbial communities which have driven progress in global knowledge of prokaryote ecology. Given the previously discussed advances in sequencing technologies sequence based metagenomics will play an increasingly important role in microbiological research. This thesis concentrates on the use of sequence data in metagenomic analyses.

8. Projects in Metagenomics: the State of the Art

Metagenomics first rose to prominence in 2001 with the discovery of light driven proteorhodopsin pumps in bacterioplankton (Béjà et al. 2001). Later, this presumably rare system was surprisingly found to be exceptionally widespread in marine prokaryotes (Sabeji et al. 2003). Venter and colleagues then applied vast effort to sequencing the microbial community of the oligotrophic Sargasso sea ecosystem near Bermuda (Venter et al. 2004). This dataset is so large it was still being analysed in recent years (Piganeau et al. 2008, Biers et al. 2009). Despite this effort, it was quickly recognised that the majority of sequences could not be assembled into larger contiguous DNA fragments as had been hoped. This study, and the follow-up Global Ocean Sampling expedition

(Rusch et al. 2007) helped to generate impetus for metagenomics and the development of the programs necessary to analyse the data. At a stroke these efforts added millions of novel putative ORFs (Yooseph et al. 2007) to the protein databases. These studies targeted bacteria and viruses, with the focus placed more on bacteria, though other researchers have specifically sampled viruses (Edwards and Rohwer 2005). Global sampling of bacteria in marine habitats has stimulated an entire genomic sequencing programme (Moore microbial programme) for marine prokaryotes.

While the vast expanse of the seas and their importance in geochemical cycling focussed much interest on marine environments, other work has targeted a range of different communities. Extremal environments are typically of interest to curious researchers (acid mine drainage; Tyson et al. 2004, whale fall degradation community; Tringe et al. 2005; indoor airborne metagenome; Tringe et al. 2008), a wide range of environmental habitats have already been initially surveyed (Dinsdale et al. 2008b, Willner et al. 2009) and clinical sampling is becoming a reality (human gut microbiomes; Gill et al. 2006, Kurokawa et al. 2007; Dinsdale et al. 2008a). Several studies have also been directed towards implicating microbial community composition in health and disease states in the mice, obese and lean humans (Gill et al. 2006, Turnbaugh et al. 2007, Turnbaugh et al. 2009). As the human associated microbiome may contain a factor of ten more genes than humans themselves carry, clinicians and scientists alike are becoming interested in the human "holobiome" (Turnbaugh et al. 2009). Furthermore, metagenomic approaches will potentially lead to discovery of novel viral and bacterial interactants in various illnesses such as Crohn's disease, psoriasis (Grice et al. 2009), hepatitis (Sorber et al. 2008) and perhaps even in the comparatively well studied microbial communities in lungs of cystic fibrosis patients. The role of anaerobes has undoubtedly been overlooked due to their resistance to culturing (Bittar et al. 2008). Metatranscriptomics, sequencing of reverse-transcribed mRNA sequences generated from environmental communities, has already been applied in multiple studies with success to assess the gene level response of a prokaryotic community to changing conditions (Gilbert et al. 2008). As such this approach continues on from the metagenomic gene level assessment of community functional potential to actually assessing which genes are transcribed at any time.

Early metagenomic analyses used a similar analysis approach to that used in single genome sequencing projects (Kunin et al. 2008). After random shotgun sequencing, reads were screened for contamination. In parallel to screening for rRNA and other established marker genes, reads were assembled into contiguous sequences (contigs) using computational assembly software (Venter et al. 2004). The aim here was to assemble large contigs, and perhaps even completely recover genomes

from unculturable environmental organisms. However, as prokaryotic diversity increases fewer reads are obtained from each organism, and contigs decrease in size. This was a problem for early taxonomic classifiers such as Tetra (Teeling et al. 2004) and Phylopythia (McHardy et al. 2007), which require large amounts of sequence information (i.e. long contigs) to attribute contigs effectively to various taxa. However, other typically Blast or Hidden Markov Model based classifiers such as Megan (Huson et al. 2007), Carma (Krause et al. 2008), Phymm (Brady and Salzberg 2009) or the SEED (Aziz et al. 2008) can work with actual sequence reads (as opposed to contigs). From a theoretical point of view using contigs to assess relative abundance of species in the environment should be discouraged, since a) different contigs will invariably have different coverages and so become incomparable and b) combining reads obscures the number of direct observations of a species (Biers et al. 2009). Researchers using the 454 pyrosequencing platform have encouragingly avoided building contigs and instead directly attributed reads on the nucleotide and protein level (Dinsdale et al. 2008a, Dinsdale et al. 2008b).

Sequencing technology will have a considerable impact on future developments in the field. For example short reads have rarely been used to date (with the exception of Sorber et al. 2008) due to lack of rapid and reliable analytical tools and their questionable utility in discovering new genes, not to mention the inherent difficulties with sequence assembly. This is likely to change with their capacity to generate large quantities of data at read lengths now approaching 100 bp. While bioinformatic tools are continually improving, the rapid generation of data will further emphasise bioinformatics as the rate-limiting step in future metagenomic studies.

9. Challenges in metagenomics

In many respects metagenomics is still in its infancy, as evidenced by the failure to date of both reference sequence collections and algorithms to adequately assess the diversity of a Minnesota soil metagenome (Tringe et al. 2005, McHardy et al. 2007). Many lessons from the field of Ecology will have to be assimilated. For example, the same microbes may play different roles in various habitats, and small changes in community structure may lead to large impacts on function. The roles of commensals and pathogens will likely blur under subtly different conditions. Community assembly rules may not be transferable from one habitat to the next. Indeed, the true scale of the problem at hand will have to be learned before systematic knowledge of microbial communities can be gathered. Ease of data generation is already beginning to exceed analytical capability, and bioinformatics is a limiting factor in the field. Already the sheer number of sequences available are

limiting the use of traditional search tools such as all against all comparisons with Blast (Hugenholtz and Tyson 2008). As metagenomics and the associated metadata on environmental parameters (Garrity et al. 2008) matures, focus is changing from assessing "the" soil metagenome to comparing and contrasting various metagenomes from similar environments (Gilbert et al. 2008, Mitra et al. 2009, Turnbaugh et al. 2009). Of course, all of these difficulties fascinatingly serve to illustrate how little is known about microbial community structure, and by extension how much remains to be discovered.

10. References

- Avery, O. T., MacLeod, C. M. & McCarty, M. (1944) Studies on the chemical nature of the substances inducing transformation of pneumococcal types. *J. Exp. Med.* 79:137-158.
- Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75.
- Beaber, J. W., Burrus, V., Hochhut, B. & Waldor, M. K. (2002) Comparison of SXT and R391, two conjugative integrating elements: definition of a genetic backbone for the mobilization of resistance determinants. *Cell Mol Life Sci* 59(12) 2065-2070.
- Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* 411(6839) 786-789.
- Biers, E. J., Sun, S. & Howard, E. C. (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* 75(7) 2221-2229.
- Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., Hampson, D. J., Bellgard, M., Wassenaar, T. M. & Ussery, D. W. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* 6(3) 165-185.
- Bittar, F., Richet, H., Dubus, J.-C., Reynaud-Gaubert, M., Stremler, N., Sarles, J., Raoult, D. & Rolain, J.-M. (2008) Molecular detection of multiple emerging pathogens in sputa from cystic fibrosis patients. *PLoS One* 3(8) e2908.
- Bohlin, J., Skjerve, E. & Ussery, D. W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* 4(4) e1000057.
- ten Bosch, J. R. & Grody, W. W. (2008) Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol Diagn* 10(6) 484-492.
- Brady, A. & Salzberg, S. L. (2009) Phymm and PhymmBL: metagenomic phylogenetic

classification with interpolated Markov models. *Nat Methods* Epub August 2.

Bush, E. C. & Lahn, B. T. (2006) The evolution of word composition in metazoan promoter sequence. *PLoS Comput Biol* **2**(11) e150.

Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**(6839) 786-789.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A. & Stratton, M. R. (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A* **105**(35) 13081-13086.

Cortez, D., Forterre, P. & Gribaldo, S. (2009) A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol* **10**(6) R65.

D'Argenio, D. A., Wu, M., Hoffman, L. R., Kulasekara, H. D., Déziel, E., Smith, E. E., Nguyen, H., et al. (2007) Growth phenotypes of *Pseudomonas aeruginosa* lasR mutants adapted to the airways of cystic fibrosis patients. *Mol Microbiol*, **64**: 512-533.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**, 629-632.

Dinsdale, E. A., Pantos, O., Smriga, S., Edwards, R. A., Angly, F., Wegley, L., Hatay, M., Hall, D., et al. (2008) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3**(2) e1584.

Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**(5) 414-424.

Edwards, R. A. & Rohwer, F. (2005) Viral metagenomics. *Nat Rev Microbiol* **3**(6) 504-510.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Jou, W. M., Molemans, F., Raeymaekers, A., den Berghe, A. V., Volckaert, G. & Ysebaert, M. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**(5551) 500-507.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & Merrick, J. M. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**(5223) 496-512.

Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tümmeler, B. & Reva, O. N. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**, 333.

Garrity, G. M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S.-A., Angiuoli, S., Cole, J. R., et al. (2008) Toward a standards-compliant genomic and metagenomic publication record. *OMICS* **12**(2) 157-160.

- Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P. & Joint, I. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* **3**(8) e3042.
- Gilbert, W. & Maxam, A. (1973) The nucleotide sequence of the *lac* operator. *Proc Natl Acad Sci U S A* **70**(12) 3581-3584.
- Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. & Nelson, K. E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**(5778) 1355-1359.
- Godde, J. S. & Bickerton, A. (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62**(6) 718-729.
- Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C. et al. (2009) Topographical and temporal diversity of the human skin microbiome. *Science* **324**(5931) 1190-1192.
- Hacker, J. & Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* **2**(5) 376-381.
- Holden, M. T. G., Feil, E. J., Lindsay, J. A., Peacock, S. J., Day, N. P. J., Enright, M. C., Foster, T. J., et al. (2004) Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A* **101**(26) 9786-9791.
- Hugenholtz, P. & Tyson, G. W. (2008) Microbiology: metagenomics. *Nature* **455**(7212) 481-483.
- Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**(3) 377-386.
- Jacob, F. & Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356.
- Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W. & Crook, D. W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* **33**(2) 376-393.
- Karlin, S., Mrázek, J. & Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12) 3899-3913.
- Klockgether, J., Würdemann, D., Reva, O., Wiehlmann, L. & Tümmler, B. (2007) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J Bacteriol* **189**(6) 2443-2459.
- Koonin, E. V. & Wolf, Y. I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* **36**(21) 6688-6719.

- Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A. & Stoye, J. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36**(7) 2230-2239.
- Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. (2008) A bioinformaticians guide to metagenomics. *Microbiol Mol Biol Rev* **72**(4) 557-78.
- Kunin, V., Sorek, R. & Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**(4) R61.
- Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**(4) 169-181.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Consortium, I. H. G. S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822) 860-921.
- Lee, D. G., Urbach, J. M., Wu, G., Liberati, N. T., Feinbaum, R. L., Miyata, S., Diggins, L. T., He, J., et al. (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol* **7**(10) R90.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057) 376-380.
- Marín, A. & Xia, X. (2008) GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. *J Theor Biol* **253**(3) 508-513.
- Maróti, G., Tong, Y., Yooseph, S., Baden-Tillson, H., Smith, H. O., Kovács, K. L., Frazier, M., Venter, J. C. & Xu, Q. (2009) Discovery of a [NiFe]-hydrogenase in metagenomic Sargasso Sea DNA: Cloning and Functional Analysis in *Thiocapsa roseopersicina*. *Appl Environ Microbiol*. Epub July 24.
- Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., Rokas, A., Yandava, C. N., et al. (2008) Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci U S A* **105**(8) 3100-3105.
- Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. (1998) "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc Natl Acad Sci U S A* **95**(7) 3943-3948.
- McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**(1) 63-72.
- Mitra, S., Klar, B. & Huson, D. H. (2009) Visual and Statistical Comparison of Metagenomes. *Bioinformatics* **25**, 1849-1855.

- Mrázek, J. (2009) Phylogenetic signals in DNA composition: limitations and prospects. *Mol Biol Evol* **26**(5) 1163-1169.
- Ochman, H., Lerat, E. & Daubin, V. (2005) Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci U S A* **102** Suppl 1, 6595-6599.
- van Passel, M., Bart, A., Luyf, A., van Kampen, A. & van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7**(1) 26.
- van Passel, M. W. J., Luyf, A. C. M., van Kampen, A. H. C., Bart, A. & van der Ende, A. (2005) Deltarho-web, an online tool to assess composition similarity of individual nucleic acid sequences. *Bioinformatics* **21**(13) 3053-3055.
- Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**(6819) 529-533.
- Piganeau, G., Desdevises, Y., Derelle, E. & Moreau, H. (2008) Picoeukaryotic sequences in the Sargasso Sea metagenome. *Genome Biol* **9**(1) R5.
- Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2) 145-158.
- Pushkarev, D., Neff, N. F. & Quake, S. R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* Epub August 10.
- Rappé, M. S. & Giovannoni, S. J. (2003) The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369-394.
- Reva, O. N. & Tümmler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**, 90.
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**(3) e77.
- Sabehi, G., Massana, R., Bielawski, J. P., Rosenberg, M., Delong, E. F. & Bèjà, O. (2003) Novel Proteorhodopsin variants from the Mediterranean and Red Seas. *Environ Microbiol* **5**(10) 842-849.
- Sanger, F. & Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**(3) 441-448.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12) 5463-5467.
- Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. (2009) RNA cytosine methylation analysis by

bisulfite sequencing. *Nucleic Acids Res* **37**(2) e12.

Sorber, K., Chiu, C., Webster, D., Dimon, M., Ruby, J. G., Hekele, A. & DeRisi, J. L. (2008) The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE* **3**(10) e3495.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P. & Rubin, E. M. (2005) Comparative metagenomics of microbial communities. *Science* **308**(5721) 554-557.

Tringe, S. G., Zhang, T., Liu, X., Yu, Y., Lee, W. H., Yap, J., Yao, F., Suan, S. T., Ing, S. K., Haynes, M., Rohwer, F., Wei, C. L., Tan, P., Bristow, J., Rubin, E. M. & Ruan, Y. (2008) The airborne metagenome in an indoor urban environment. *PLoS ONE* **3**(4) e1862.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. (2007) The human microbiome project. *Nature* **449**(7164) 804-810.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R. & Gordon, J. I. (2009) A core gut microbiome in obese and lean twins. *Nature* **457**(7228) 480-484.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. & Banfield, J. F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**(6978) 37-43.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O. et al. (2001) The sequence of the human genome. *Science* **291**(5507) 1304-1351.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**(5667) 66-74.

Vernikos, G. S., Thomson, N. R. & Parkhill, J. (2007) Genetic flux over time in the *Salmonella* lineage. *Genome Biol* **8**(6) R100.

Watson, J. D. & Crick, F. H. (1953) Molecular structure of nucleic acids, a structure for deoxyribose nucleic acid. *Nature* **171**(4356) 737-738.

Willner, D., Thurber, R. V. & Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol.* **11**(7) 1752-1766.

Worning, P., Jensen, L. J., Hallin, P. F., Staerfeldt, H.-H. & Ussery, D. W. (2006) Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* **8**(2) 353-361.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**(3) e16.

Objectives

The major goals of this thesis were to investigate the distributions of mid-length oligomers and their potential utility in both genomic and metagenomic analysis. Oligomer analysis is becoming increasingly useful in rapid analysis of sequence data, which is now almost routine. Past work has mainly centred on the use of shorter oligomers up to 7 base pairs in length in applications such as development of genome signatures, differentiation of foreign DNA in a genome, and testing the suitability of whole genome signatures for phylogeny.

Work was mainly carried out at Hanover Medical School with extensive collaboration with Dr. Oleg Reva in Pretoria, South Africa.

This thesis presents analyses of the following topics:

1. Visualisation and segmentation of genomes according to tetramer usage parameters. Data are presented in an online Genome Browser.
2. Comparative assessments of the genus *Pseudomonas* and phylum Chlorobi with 4-14 bp oligonucleotides. Overrepresented DNA sequences and peptides were implicated and potential for their usage in metagenomic analyses were suggested. Whole proteome comparisons of 11 Chlorobi strains were also performed and discussed in a phylogenetic context.
3. Discovery of a small group of abundant oligomers present in the majority of sequenced bacterial lineages.
4. Comparative analysis of global oligonucleotide distributions across all sequenced bacteria.
5. Microevolution of *Pseudomonas aeruginosa* PAO1 sublines in the laboratory, following Illumina GA20 sequencing, reference and *de novo* assembly.
6. Assessment of the taxonomic fidelity of short 36 bp sequencing reads and their ability to specifically and reliably discern between bacterial taxa at various taxonomic levels.

We have now established and made available two alternative analysis and visualisation systems for oligonucleotide usage in bacterial genomes. These methods were successfully tested within sequencing projects on several *Pseudomonas aeruginosa* strains. Furthermore, due to the research need generated by the growing impact of modern nucleic acid sequencing on scientific research, a fast and sensitive system for analysis of short metagenomic reads was implemented. This system is particularly promising for the future expansion of sequencing into fast and accurate monitoring of strains within biological communities.

Results

1. Comparative genomics of the genus *Pseudomonas* and phylum Chlorobi

1.1 Background

The genus *Pseudomonas* is a clade of ubiquitous, free living gram-negative gammaproteobacteria which are of medical (*P. aeruginosa*), environmental (*P. putida*, *P. fluorescens*, *P. syringae*), and biotechnological (*P. putida*) significance. Pseudomonads are successful in so many different environments due to their metabolic and physiological diversity conferred by their large genomes (dos Santos et al. 2004). *P. aeruginosa*, for example, has even been observed growing in bottles of hospital detergents (B. Tümmler, *pers. comm.*). Previous work included the sequencing, assembly and annotation of *Pseudomonas putida* KT2440 as part of an international team (Nelson et al. 2002). While strains within several species have been compared against one another (Lee et al. 2006, Silby et al. 2009), no study had yet assessed the genomic variability of multiple *Pseudomonas* species. Following the completion of *P. aeruginosa* PA7 in 2007 thirteen genomes from the genus were available for comparative bioinformatic analysis. This made *Pseudomonas* one of the genera best represented by whole genome sequences at the time, and a good system for comparative genomics.

The Chlorobi, or green sulfur bacteria, are photosynthetic anoxygenic phototrophs which are abundant in anaerobic low-light environments. Their curious lifestyles are reflected in the sampling sites from which they were discovered. For example, the first Chlorobi species sampled, *Chlorobaculum* (= *Chlorobium*) *tepidum* TLS, was isolated from a high-sulfide acidic hot spring in New Zealand (Eisen et al. 2002). In the phylum Chlorobi, many complete genome sequences have been published in the last few years bringing the total to 11. However no general comparative genomics paper has to our knowledge yet appeared on these interesting bacteria, as interest has chiefly been focused on the evolution of the photosynthetic apparatus. These two manuscripts were intended to shed light on the multiple genomes of these two phylogenetic groups using current comparative genomic techniques.

1.1.1 Comparative genomics

Long prior to the appearance of sequenced genomes several groups were already working on comparative analysis of microbial genomes. These early comparators used biochemical assessment of GC content to ascertain the possible range of observed values in microbial species. A range of GC from approximately 25 % to 75 % had already been observed in 1962 (Sueoka 1962). In the

pre-genomic era researchers could not make extensive predictions or surveys of the metabolic or functional capacity of microbes. Comparative analyses gained in usefulness following the accrument of genomic data after 1995.

Comparative genomics leverages the data provided by the growing number of completed genome sequences to provide information on prokaryote phylogeny, evolutionary mechanisms such as quantitative assessment of selection pressure, and genotype/phenotype correlations with pathogenicity.

Classical comparative genomics involves assessment of genes, and as more sequence data later became available, entire genomes and genomic regions. However, while gene based comparisons are highly informative, sole use of single genes for comparing prokaryotic species can be misleading given the degree of horizontal gene transfer known to have taken place. Incidentally, recognition of the degree of horizontal genetic exchange between bacteria is one of the key achievements of comparative genomics (Ochman et al. 2005). While ribosomal RNA has been a very successful approach to resolving many different phylogenies (Olsen and Woese 1993) and is still the most widely used phylogenetic marker gene, it is not always sufficient for discerning between two strains. Complementary analyses which can take into account information from the whole genome would be advantageous, and oligonucleotide signatures are one of these.

1.1.2 Oligonucleotide signatures

Oligonucleotide patterns are statistical parameters based on normalised counts of short oligomers in DNA sequences (Karlin et al. 1997). It has been observed that the distribution of mono- and oligonucleotide patterns are relatively homogenous within most regions of bacterial chromosomes, yet differ markedly between various taxa (Karlin et al. 1997). In other words, they carry a phylogenetic signal (Pride et al. 2003, Willner et al. 2009). This fact potentially confers oligomers a greater utility than simple GC content in a range of applications due to their greater variation. There must be a range of physico-chemical factors that act on the DNA sequence to restrict variation to certain constraints across the entire genome sequence (Reva and Tümmler 2004). Indeed, a range of environment associated factors such as temperature, salinity, and radiation have been demonstrated to exert an influence on nucleotide composition (Foerstner et al. 2005). Furthermore, intrinsic factors such as replication and repair machinery, restriction modification systems, codon bias and structural parameters (Karlin et al. 1997, Pride et al. 2003, Reva and Tümmler 2004, Bohlin et al. 2008) must play a role.

Karlin and coworkers introduced the concept of dinucleotides delta-difference, a measure for discerning within and between two nucleotide sequences (Karlin et al. 1997). While dinucleotides, of which only 16 combinations are possible, provide reliable statistics using just short DNA sequences (Reva and Tümmler 2004, Willner et al. 2009), larger oligomers provided several advantages such as greater sensitivity and specificity when comparing large numbers of genomes if sequence length is not limiting. Subsequent work revealed tetramers are an optimal tradeoff between information content and sequence size required for reliable statistics (Becq et al. 2007), and to exceed the signal from codon bias (Pride et al. 2003). Phylogenetic trees based on dinucleotide patterns are less similar to the "gold standard" rRNA trees than those based on tetramer zero-order Markov models (Pride et al. 2003). In other words, tetranucleotides better reflect bacterial phylogeny than dinucleotides (Teeling et al. 2004), and furthermore not controlling for biases by using a simple zero-order Markov model instead of a more complicated system contributes to a clearer phylogenetic signal (Pride et al. 2003). This fact further strengthens the case for using higher order oligomers for discrimination between genomes. A further reason for studying intermediate length 8-14 bp oligomers is that most analyses, for example Bush and Lahn (2006), avoided investigating oligomer with sizes larger than octamers.

Other groups introduced further oligonucleotide based parameters into the literature. These include Distance, Pattern skew and Oligonucleotide variance (Reva and Tümmler 2004, 2005). Distance is the a mathematical measure of divergence between two patterns of oligonucleotide frequencies, expressed as a percentage of the maximum possible mathematical distance. This is typically the mathematical distance between a pattern calculated for a local pattern versus that calculated for a whole genome. Pattern skew is the distance between oligomer patterns on the direct and reverse DNA strands of the same nucleotide sequence. Lastly, Oligonucleotide variance is the variance of oligonucleotides, where a lower value indicating that the oligomer usage in a region is more restricted, such as in repeat regions. Previous work had been focussed on using oligonucleotide signatures to compare *P. aeruginosa* and *P. putida* (Weinel et al. 2002a), and this approach was continued and extended to longer oligomers and more genomes in this thesis.

1.2 About the manuscripts

These two manuscripts were grouped together due to their common ground in comparative genomics. It was the intention of the first paper to i) attempt to locate overrepresented 8-14 bp oligomers in *Pseudomonas* strains, ii) analyse genomic structure including variable regions and

genomic islands using previously investigated tetranucleotides, and compare and contrast these to mid-length oligomers and iii) use *Pseudomonas* as a test case for locating, visualising and testing specificity of new oligomeric characters in metagenomics. The major results from this paper included creation of the OligoCounter program suite, which formed the core tools for future work within this thesis. A dataset of overrepresented oligomers could be generated and maintained with this suite. The discovery by Weinel et al. (2002b) of 8-14 bp oligomers in *Pseudomonas* strains was verified and extended to newly sequenced genomes. Visualisation of whole genomes demonstrated usage of 8-14 bp oligonucleotides to reflect that of the shorter tetramers, which was surprisingly in contrast to past predictions (Reva and Tümmler 2005). Regions of atypical genomic oligomer usage, such as genomic islands, tend to lack globally overrepresented 8-14mers. This can be explained by the fact genomic islands and phages also exhibit their own characteristic oligonucleotide usage or be similar to a putative previous host genome (Pride et al. 2003). In other words, divergent genomic regions are frequently involved in horizontal gene transfer and this process is a major confounder of phylogenetic and taxonomic inference. Thus, oligomers appeared in theory to be promising candidates for taxonomic markers in the field of metagenomics. In practice, overrepresented oligomers were located in all *Pseudomonas* strains in sufficient numbers, even at a very conservative overrepresentation cutoff, and displayed significant specificity to allow distinction between species. Another more anecdotal observation concerned the apparent overrepresentation of a few tripeptides in *Pseudomonas*. The amino acid leucine was notably present in the triplets despite normalisation for its expected high frequency.

The second manuscript, "Comparative genomics of the Chlorobi", also makes use of oligonucleotide signatures as well as a range of other bioinformatic techniques in an assessment of the eleven sequenced genomes of the green sulfur bacteria. Oligonucleotide parameters precisely and elegantly depict the genome positions of a number of giant genes with constrained amino acid usage (see Figure 1 in the manuscript), but also show genomic islands, islets and phages. The phylogeny of the Chlorobi have been the subject of considerable debate (Imhoff 2003) and is likely a model phylum for the positive impact of molecular taxonomy over traditional morphological characteristics. Our findings, based on the comparisons of the sequences of two appropriate genes and the entire proteome content for all completely sequenced genomes, support those of Imhoff (2003). Using a whole proteome sequence comparison approach we also noted the conservation of orthologous genes throughout the Chlorobi, although gene synteny was, as expected, not conserved in this phylum. The origin of photosynthesis and horizontal gene transfer events of various

photosynthetic genes have been of great interest in studies of these taxa. Our results of putatively horizontally transferred regions infer that, while some metabolic genes associated with photosynthesis have apparently been transferred between taxa, several hundred genes are probably necessary for the demanding requirements of a photosynthetic lifestyle. Thus most of the genes which enable photosynthesis belong to the core genome.

The first paper, "Visualization of Pseudomonas genomic structure by abundant 8-14mer oligonucleotides" was published in Environmental Microbiology in 2009 while "Comparative genomics of the Chlorobi" had been submitted to Photosynthesis Research at the time of writing. I was lead author on both works, and contributed written passages, figures, programming, and analyses to both. Oleg Reva contributed analyses and figures to the first paper, while David Ussery contributed whole genome proteome assessment figures and commented on the second manuscript. Burkhard Tümmeler wrote parts of both works.

1.3 References

Becq, J., Gutierrez, M. C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O. & Deschavanne, P. (2007) Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol Biol Evol* **24**(8) 1861-1871.

Bohlin, J., Skjerve, E. & Ussery, D. W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* **4**(4) e1000057.

Bush, E. C. & Lahn, B. T. (2006) The evolution of word composition in metazoan promoter sequence. *PLoS Comput Biol* **2**(11) e150.

Eisen, J. A., Nelson, K. E., Paulsen, I. T., Heidelberg, J. F., Wu, M., Dodson, R. J., Deboy, R., Gwinn, M. L., et al. (2002) The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proc Natl Acad Sci U S A* **99**(14) 9509-9514.

Foerstner, K. U., von Mering, C., Hooper, S. D. & Bork, P. (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep* **6**(12) 1208-1213.

Imhoff, J. F. (2003) Phylogenetic taxonomy of the family Chlorobiaceae on the basis of 16S rRNA and *fmo* (Fenna-Matthews-Olson protein) gene sequences. *Int J Syst Evol Microbiol* **53**(Pt 4) 941-951.

Karlin, S., Mrázek, J. & Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12) 3899-3913.

- Lee, D. G., Urbach, J. M., Wu, G., Liberati, N. T., Feinbaum, R. L., Miyata, S., Diggins, L. T., He, J., et al. (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol* **7**(10) R90.
- Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., dos Santos, V. A. P. M., Fouts, D. E., et al. (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* **4**(12) 799-808.
- Ochman, H., Lerat, E. & Daubin, V. (2005) Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci U S A* **102** Suppl 1, 6595-6599.
- Olsen, G. J. & Woese, C. R. (1993) Ribosomal RNA: a key to phylogeny. *FASEB J* **7**(1) 113-123.
- Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2) 145-158.
- Reva, O. N. & Tümmler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**, 90.
- Reva, O. N. & Tümmler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.
- Santos, V. A. P. M. D., Heim, S., Moore, E. R. B., Strätz, M. & Timmis, K. N. (2004) Insights into the genomic basis of niche specificity of *Pseudomonas putida* KT2440. *Environ Microbiol* **6**(12) 1264-1286.
- Silby, M. W., Cerdeño-Tárraga, A. M., Vernikos, G. S., Giddens, S. R., Jackson, R. W., Preston, G. M., et al. (2009) Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. *Genome Biol* **10**(5) R51.
- Sueoka N. (1962) On genetic basis for variation and heterogeneity of DNA base composition. *Proc. Natl. Sci. USA* **48**, 582-92.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163.
- Weinel, C., Nelson, K. E. & Tümmler, B. (2002a) Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ Microbiol* **4**(12) 809-818.
- Weinel, C., Ussery, D. W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C. & Tümmler, B. (2002b) Comparative Genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: Orthologs, Codon Usage, Repetitive Extragenic Palindromic Elements, and Oligonucleotide Motif Signatures. *Genome Letters* **4**, 175-187.
- Willner, D., Thurber, R. V. & Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol.* **11**(7) 1752-1766.

Visualization of *Pseudomonas* genomic structure by abundant 8–14mer oligonucleotides

Colin F. Davenport,^{1*} Lutz Wiehlmann,¹

Oleg N. Reva² and Burkhard Tümmler¹

¹Klinische Forschergruppe, OE 6711, Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, D-30625 Hanover, Germany.

²Biochemistry Department, Bioinformatics and Computational Biology Unit, University of Pretoria, Hillcrest, Pretoria 0002, South Africa.

Summary

Under- and over-represented mono- to hexanucleotides are signatures of bacterial genomes, but the compositional biases of octa- to tetradecanucleotides have not yet been explored. Thirteen completely sequenced genomes of the *Pseudomonas* genus were searched for highly overrepresented 8–14mers. Between 59–989 overrepresented 8–14mers were found to exceed the applied threshold value. All genomic data sets of the 13 strains showed a consistent pattern, with individual oligomers clustering in either non-coding or coding regions. Non-coding oligonucleotides were typically part of longer repeats. Coding oligonucleotides were evenly distributed in the core genome, preferred one reading frame and matched with the local tetranucleotide usage patterns. Genomic islands were recognized by the depletion of overrepresented oligonucleotides. Several mainly coding 8–14mers occurred in genomes on average every 10 000 bp or less. Such frequently occurring 8–14mers could become useful markers for species identification. In the future of next-generation ultra-high throughput DNA sequencing, the composition of bacterial metagenomes may be quantified by scanning the primary sequence reads for these 8–14mer markers.

Introduction

Bacterial genome sequences, like the text of a book, can be analysed in terms of their meaning or their grammar. The majority of readers will be familiar with the first task, i.e. the interpretation of the encoded function of a

sequence (Médigue and Moszer, 2007), but less familiar with the second task, i.e. the identification of compositional rules that govern genome organization and structure (Allen *et al.*, 2006). The initial studies in the latter field focused on short mono- to hexanucleotides that are either under- or over-represented in a bacterial genome (Burge *et al.*, 1992; Karlin *et al.*, 1997). The compositional biases of these short oligonucleotide words were found to represent genomic signatures. Frequencies of the short oligonucleotides depend on their physicochemical properties such as base-stacking energies, propeller twist or bendability (Baldi and Baisnee, 2000), codon usage bias (Sharp and Li, 1987; Willenbrock *et al.*, 2006) and biological constraints of, e.g. DNA replication and repair, transcriptional or translational control, or restriction-modification systems (Gelfand and Koonin, 1997; Karlin *et al.*, 1997).

Mono- to hexanucleotide usage in a bacterial genome is known to contain a phylogenetic signal (Pride *et al.*, 2003; Reva and Tümmler, 2004; Teeling *et al.*, 2004; Bohlin *et al.*, 2008a,b; Reva *et al.*, 2008). Thus, it should be principally possible to exploit oligonucleotide usage (OU) patterns for the identification of bacterial species. An individual short oligonucleotide word, however, is inherently unspecific and consequently one needs a long stretch of contiguous DNA sequence for a reliable taxonomic classification (Reva *et al.*, 2008). Thus, OU pattern analysis may *prima facie* not be superior to common 16S rDNA sequencing (Olsen and Woese, 1993). This disadvantage, however, may be overcome, if one could extract information from longer, more specific oligonucleotide words that are sufficiently frequent to occur in a sample of shotgun sequences. Hence, we hypothesized that a data set of highly overrepresented oligomers of intermediate size, i.e. 8–14mers, could be useful for species identification even in heterogeneous samples derived from microbial communities. These sequences that occur by chance every 65 kilobases (kb) (8mers) to 268 megabases (14mers) should be found at least once every 50 kb to allow a meaningful analysis of short reads. The new next-generation sequencing technologies (Schuster, 2008) can generate several hundred megabases of primary sequence within a single run. Sequence reads are short, yet frequently long enough to contain oligonucleotide words. Hence, searching the primary sequence data for overrepresented oligonucleotides of intermediate size

Received 23 July, 2008; accepted 3 November, 2008. *For correspondence. E-mail davenport.colin@mh-hannover.de, Tel. (+49) 511 532 6721; Fax (+49) 511 532 6723.

Table 1. Overrepresented octa- to tetranucleotides in the genus *Pseudomonas*.

Strain	% coding	No. of oligomers	Median total occurrences of one oligomer	Median occurrences thereof in coding regions (%)	Median average distance (bp) between occurrences of the same oligomers
<i>P. aeruginosa</i> PAO1	89	485	623 (323–1061)	97.7 (96.8–98.5)	10 039 (5901–19349)
<i>P. aeruginosa</i> PA7	89	527	640 (329–1120)	98.4 (97.5–98.9)	10 278 (5876–19995)
<i>P. aeruginosa</i> PA14	89	452	640 (350–1108)	97.9 (97.1–98.6)	10 207 (5894–18625)
<i>P. entomophila</i> L48	88	989	164 (88–348)	3.8 (1.8–29.8)	35 689 (16861–66166)
<i>P. fluorescens</i> PfO-1	89	92	727 (423–986)	98.1 (97.4–98.7)	8 863 (6519–15202)
<i>P. fluorescens</i> Pf-5	88	807	275 (155–683)	97.3 (96.4–98.4)	25 633 (10351–45279)
<i>P. mendocina</i> ymp	89	483	398 (158–755)	97.6 (96.4–98.4)	12 713 (6712–31854)
<i>P. putida</i> F1	88	458	160 (86–227)	4.5 (3.1–10.4)	37 018 (26140–68505)
<i>P. putida</i> KT2440	86	310	147 (104–273)	10.1 (8.2–94.1)	41 769 (22561–58874)
<i>P. stutzeri</i> A1501	89	133	618 (346–893)	96.5 (95.9–97.2)	7 378 (5108–13181)
<i>P. syringae</i> DC3000	85	59	365 (94–819)	90.1 (81.5–94.9)	17 478 (7800–67338)
<i>P. syringae</i> B728a	87	198	101 (87–123)	26.6 (12.3–30.6)	60 038 (49142–69246)
<i>P. syringae</i> 1448a	84	97	111 (87–185)	12.8 (7.8–16.8)	52 935 (31875–67372)

Total occurrences in the genome and median distance were determined for each oligomer. The distribution of the individual values is presented for the whole data set of oligomers in columns 4, 5 and 6 as medians and inner quartiles.

could be a promising approach to determine the presence of individual taxa in a bacterial metagenome.

With these potential future application for bacterial metagenomics in mind, we hypothesized (i) oligomers in the 8–14 bp size range exist that are overrepresented over and above what would be normally expected in prokaryotic genomes; (ii) these oligomer distributions will reveal information about genomic structure; and (iii) widespread oligomers will be suitable as markers to identify various taxa.

To test these hypotheses, 13 completely sequenced genomes of the gamma-proteobacterial genus *Pseudomonas* (*sensu strictu*) were analysed in their contents of highly overrepresented octa- to tetradecanucleotides. These ubiquitous bacteria were chosen because they enjoy a versatile lifestyle and are capable of colonizing a wide range of niches that are targets for metagenomics, including soil and aquatic habitats, animal tissues, the plant rhizosphere and phyllosphere (Ramos, 2004). Highly overrepresented oligonucleotide words were detected in all studied *Pseudomonas* genomes. Each genome data set was found to consist of two distinct, non-overlapping classes of words. One class of 8–14mers clusters in non-coding repeats. The other class of 8–14mers is evenly distributed in coding regions of the core genome (Mathee *et al.*, 2008), and each oligomer shows strong preponderance of one coding frame. The latter class contains numerous markers with an average occurrence of 10 kb or less that are potentially useful for species identification.

Results

Overrepresented 8–14 bp oligomers in *Pseudomonas* genomes

Overrepresented octa- to tetradecanucleotides were found in all of the 13 analysed completely sequenced

Pseudomonas strains at the stringent χ^2 -threshold of 3000 (Table 1; Fig. 1). With a random genome of 50% G+C content, no overrepresented oligos were found at the same level of overrepresentation. Indeed, the most highly overrepresented oligos were octamers exhibiting χ^2 -values of just 20. This clearly demonstrates the significance of the overrepresented oligomers we have located in real genomes and gives some hint as to the selective pressure DNA is under for maintenance of sequence DNA structure, coding regions, binding sites and so forth.

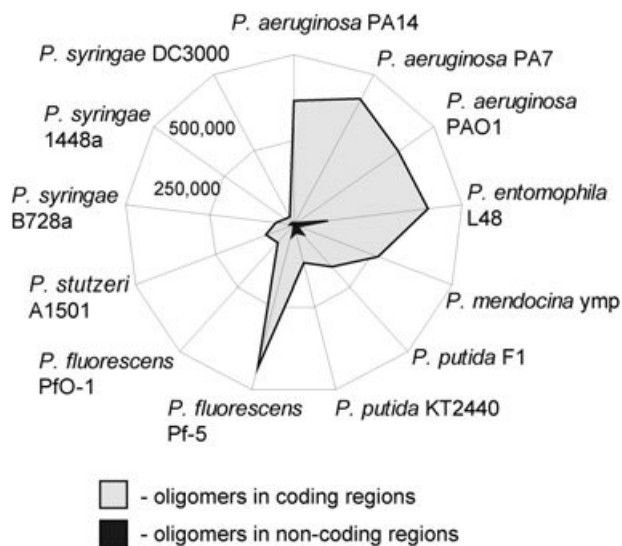


Fig. 1. Absolute counts of overrepresented octa- to tetradecanucleotides in 13 completely sequenced genomes of the *Pseudomonas* genus. The threshold for including an oligomer was a minimum count of 70 and χ^2 of 3000 or more. Counts in the coding (grey) and non-coding regions (black) are differentiated by shade.

Table 2. Strain, species and genus specificity of the over-represented octa- to tetradecanucleotides.

Strain	Strain level markers ^a	Species level markers ^a	Specific species level markers ^b	Genus level markers ^a
<i>P. aeruginosa</i> PAO1	239	199	16	7 ^c
<i>P. aeruginosa</i> PA7	252			
<i>P. aeruginosa</i> PA14	222			
<i>P. entomophila</i> L48	134	134	21	
<i>P. fluorescens</i> PfO-1	51	40	2	
<i>P. fluorescens</i> Pf-5	191			
<i>P. mendocina</i> ymp	202	202	4	
<i>P. putida</i> F1	57	53	2	
<i>P. putida</i> KT2440	55			
<i>P. stutzeri</i> A1501	95	95	0	
<i>P. syringae</i> DC3000	26	13	0	
<i>P. syringae</i> B728a	25			
<i>P. syringae</i> 1448a	14			

a. A marker occurs at an average distance of 10 000 bp or less in the genome.

b. Specific species level markers are the markers present in all strains of that species that are not overrepresented in any non-*Pseudomonas* sp. strain, which yielded data ($n = 407$) in our chi-squared 3000 data set.

c. The words are CAGCAGCA, CCAGCAGC, CGATCAGC, GCTGATCG, GCTGCTGG, TGCTGCTG, TGGCGCTG.

The applied conservative thresholds of a minimum count of 70 oligomer occurrences per genome and χ^2 of 3000 left between 59 and 989 overrepresented oligomers for each strain (Tables 1 and 2). *Pseudomonas entomophila* L48 and *P. fluorescens* Pf-5 contained the most oligomers (989 and 807) while *P. syringae* DC3000 and *P. fluorescens* PFO-1 contained the fewest (59 and 92). Differences in oligomers between the three *Pseudomonas aeruginosa*

strains, PAO1, PA7 and PA14, were far less marked. The *P. aeruginosa* strains also displayed, together with both *P. fluorescens* strains, *P. mendocina* ymp and *P. stutzeri* A1501, the highest proportions of coding oligomer occurrences (98%, Table 1). These strains also tended to have the greatest median total occurrences. This correlation can be explained in that overrepresented oligomers in coding regions tend to be present in higher numbers than longer, non-coding repeats. The strains with the lowest median percentage of coding oligomers were *P. entomophila* L48 and *P. putida* F1. Both these strains are dominated by non-coding repeat regions. Most other strains have more than 90% of their overrepresented oligomer occurrences in coding regions.

The number of marker oligomers, defined as those oligomers present every 10000 bp or less (average value) in the genome, was highest in *P. aeruginosa* PAO1 and PA14 (Table 1). The fewest markers were found in the three *P. syringae* strains. Core species level oligomer markers present in all strains of the respective species were derived. The specificity of these oligomers was tested by searching in all non-*Pseudomonas* bacterial chromosomes, with results presented as in Fig. 2, Table 2 and Table S1. Ideally, oligomers intended to be used as markers should be overrepresented in as few other taxa as possible, and thus the histograms should be skewed towards the left. *Pseudomonas aeruginosa* is the most promising candidate, with 16 markers not overrepresented in any other sequenced bacterium. Given 100-bp-long reads, a marker present every 10 000 bp would be expected in 1% of reads (100/10 000). Assuming a combination of markers is necessary to avoid false positives, perhaps around 10% of a data set could be

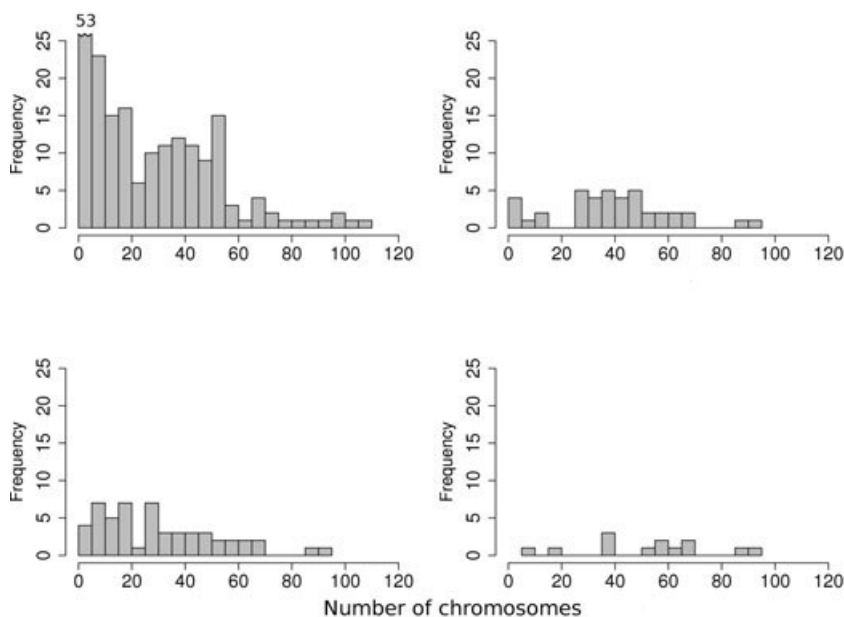


Fig. 2. Global abundance of octa- to tetradecanucleotides that are highly overrepresented in *P. aeruginosa* (top left), *P. fluorescens* (top right), *P. putida* (bottom left) and *P. syringae* chromosomes (bottom right). The oligomer markers found in the *Pseudomonas* chromosomes were scanned for their overrepresentation in completely sequenced non-*Pseudomonas* bacterial genomes. The histograms display the counts of oligomer markers which are also overrepresented in the respective number of non-*Pseudomonas* chromosomes. A skewed distribution towards the left indicates greater genomic specificity (i.e. occurrence in few chromosomes other than *Pseudomonas* sp.) of the oligomers, and a greater potential utility in identification of this species.

Table 3. Preferential occurrence of overrepresented octa- to tetradecanucleotides in either the coding or non-coding regions of genomes of the *Pseudomonas* genus.

Genome	Occurrence of overrepresented 8–14mers			Total
	Biased ($P < 0.001$)		Unbiased	
	Coding	Non-coding		
<i>P. aeruginosa</i> PAO1	459	0	26	485
<i>P. aeruginosa</i> PA7	500	0	27	527
<i>P. aeruginosa</i> PA14	432	0	20	452
<i>P. entomophila</i> L48	214	764	11	989
<i>P. fluorescens</i> PfO-1	91	0	1	92
<i>P. fluorescens</i> Pf-5	523	252	32	807
<i>P. mendocina</i> ymp	385	78	20	483
<i>P. putida</i> F1	96	361	1	458
<i>P. putida</i> KT2440	87	223	0	310
<i>P. stutzeri</i> A1501	131	0	2	133
<i>P. syringae</i> DC3000	30	29	0	59
<i>P. syringae</i> B728a	31	167	0	198
<i>P. syringae</i> 1448a	20	77	0	97

Unbiased oligomers were not significantly biased towards coding or non-coding regions.

attributed to taxa that contain specific markers. *Pseudomonas aeruginosa* rather than *P. syringae* would be a better initial candidate for these sorts of analyses given the number of markers in Table 2. As these data merely present the tip of the iceberg and a correlation between the number of markers and number of specific markers is apparent, future practical applications such as metagenome analysis may use an extended set of markers gained from more relaxed χ^2 -thresholds.

97.2% of the 5090 overrepresented oligonucleotides in the 13 genomes were significantly more frequently (χ^2 , $P < 0.001$) found in either the coding (2999 oligomers; 58.9%) or non-coding regions (1951 oligomers; 38.3%) (Table 3). Just 140 oligomers (2.8%) were randomly distributed, showing little bias towards either coding or non-coding regions. Thus there was a strong bias in all analysed genomes for having a particular 8–14mer in either coding or non-coding segments.

Genomic images provide a striking graphical summary of the distribution of overrepresented oligomers in a genome. Figure 3 shows the map positions of the oligomers sorted by decreasing χ^2 -values for six sequenced *Pseudomonads*. Most oligonucleotides of the coding regions (green) were evenly distributed in the chromosome. Overrepresented oligonucleotides in non-coding regions (black), however, typically clustered to a few map positions. These oligomers often appear as black horizontal lines, indicating that the oligonucleotides are in fact part of longer words. Related oligomers, for example 14 bp components of a longer repeat, tend to be placed together as OligoCounter considers a single nucleotide

difference as a completely separate oligomer. The same pattern of such a line typically will occur at various genome positions, implying that the long oligonucleotide represents a repeat. A well-known example is the repetitive extragenic palindromic regions (REPs) of *P. putida* KT2440 (Aranda-Olmedo *et al.*, 2002; Weinel *et al.*, 2002a) (Fig. 3). The individual long oligomers that make up a repeat occur far more frequently than would be expected, so produce exceptionally high χ^2 -values and thus occur on the left hand side of the graphics. Shorter oligomers that are typically present in much higher copy numbers occur afterwards (to the right in the figures), as they tend towards lower χ^2 -values due to their length. These shorter sequences were usually found in coding sequences in most genomes.

Figure 3 visualizes the strikingly different localization of the overrepresented oligomers in the sequenced *Pseudomonas* genomes. In *P. entomophila* L48 almost all displayed oligomers are part of non-coding repeats whereas in *P. aeruginosa* almost all oligomers encode peptides. *Pseudomonas fluorescens* and *P. putida* showed intermediate patterns.

The absence of highly overrepresented oligomers in certain genomic regions is an indicator that these areas diverge from the genomic norm. *Pseudomonas fluorescens* Pf-5 and *P. putida* KT2440 contain many divergent regions in accordance with their known mosaic structure (Weinel *et al.*, 2002b), while *P. aeruginosa* PAO1 and PA14 have very few (Fig. 3). Most divergent regions belong to the flexible accessory genome predominantly made up of transposons, phages and tRNA-integrated genomic islands (Table S2). A further category lacking the overrepresented oligomers are genomic regions that encode surface adhesion proteins (Reva and Tümmler, 2008) and various transporters. Although these genes contain many local repeats, they are distinct from the genome-wide highly overrepresented oligonucleotides (Table S2). A third category with an atypical oligonucleotide composition are the phylogenetically ancient ribosomal RNA operons (Reva and Tümmler, 2005).

Repetitive extragenic palindromic regions were identified in over half of the sequenced *Pseudomonas* strains in vastly differing copy numbers and lengths. We did not search specifically for REPs, rather they are a significant byproduct of our methodology. High REP copy numbers can be clearly seen as extensive non-coding repeats in *P. putida* KT2440 (Aranda-Olmedo *et al.*, 2002; Weinel *et al.*, 2002a) and *P. entomophila* L48 (Vodovar *et al.*, 2006) (Fig. 3). Repetitive extragenic palindromic regions were, however, not detected in *P. aeruginosa*, although they do exist (Weinel *et al.*, 2002a). They escaped notice because of the larger sequence diversity of these REPs (Weinel *et al.*, 2002a).

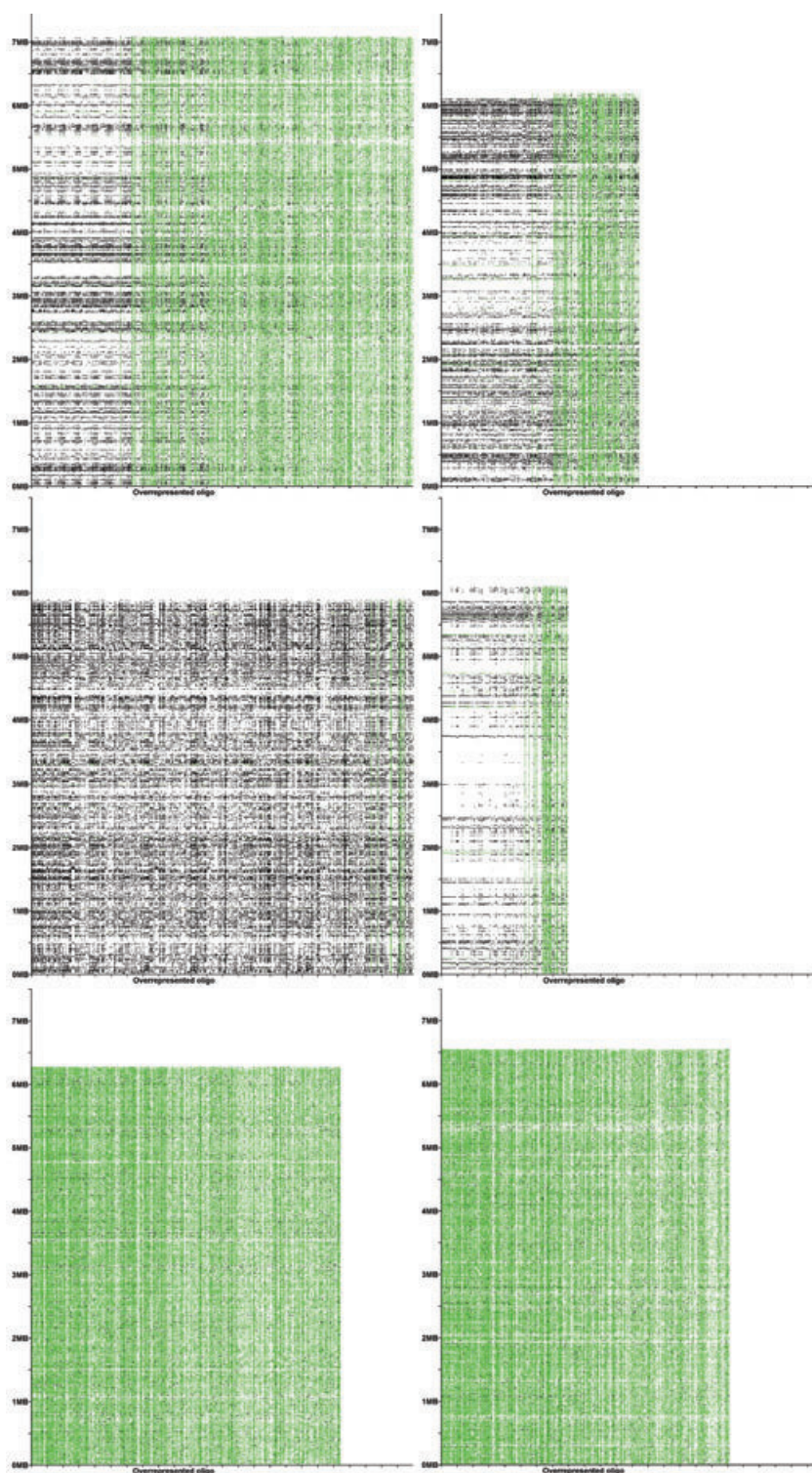


Fig. 3. The most overrepresented 8–14 bp oligomers in six strains (top left: *P. fluorescens* Pf-5; top right: *P. putida* KT2440; middle left: *P. entomophila* L48; middle right: *P. syringae* B728a; bottom left: *P. aeruginosa* PAO1 and bottom right: *P. aeruginosa* PA14) sorted by decreasing χ^2 -values. The genomic position of each oligo is plotted on the y-axis. Black or green dots are printed where an oligonucleotide occurs a in non-coding or coding region respectively. Table S2 describes the observed features. Figures were created with the program OligoViz.

Whole-genome comparisons of G+C contents, tetramer usage and distribution of overrepresented octa – to tetradecanucleotides

A genome atlas (Pedersen *et al.*, 2000; Skovgaard *et al.*,

2002) allows clear and concise presentation of a whole range of oligonucleotide parameters. Values are calculated for a selected sliding window and compared with the genomic average. As reference we chose the G+C contents and three diagnostic parameters of tetranucleotide

usage, i.e. distance, pattern skew and tetranucleotide variance (Reva and Tümmler, 2004; 2005). The third and second outermost circles of a typical genome atlas show the frequency distribution of overrepresented ($\chi^2 > 3000$) and highly overrepresented ($\chi^2 > 7000$) 8–14mers in the bacterial chromosome. The outermost ring visualizes differences between tetranucleotide usage and the frequency of the overrepresented longer oligomers. Figure 4 displays genome atlases of *P. putida* KT2440 and *P. aeruginosa* PA14. For genome atlases of the other sequenced *Pseudomonas* strains and several *Bacillus* strains for comparison, the reader is referred to Fig. S2.

All genome atlases share some common features. First, and most importantly, tetranucleotide usage and abundance of genome-wide overrepresented 8–14mers are strongly correlated with each other implying that overrepresented octa- to tetradecamers are diagnostic for the core genome and can be exploited as a genomic signature. Only very few genomic regions do not adhere to this rule (see below for some examples). Second, genomic regions with low G+C contents in the GC-rich *Pseudomonads* are characterized by high distance and pattern skew, low oligonucleotide variance and are strongly depleted in common longer oligonucleotide words. These patterns are more variable for the AT-rich *Bacillus* strains. In other words, anomalous genomic regions consistently show anomalous OU patterns for mono- to tetradecanucleotides.

Besides these common global features, the genome-specific characteristics can be extracted from the individual atlas at any resolution from a single 10 kb window to the whole chromosome as exemplified in Fig. 4 for *P. aeruginosa* PA14 and *P. putida* KT2440.

P. putida KT2440 genome atlas

Pseudomonas putida contains a mosaic genome with many divergent regions (Weinel *et al.*, 2002b; Dos Santos *et al.*, 2004). Because of *P. putida*'s relatively high G+C content (61%), most atypical regions have a lower G+C content than the average. The genome atlas displays segments at 0.6, 2.0, 2.2, 3.5 Mb, the segment between 4.2 and 5.1 Mb and the area around the origin of replication as the regions with the most divergent oligonucleotide composition in the KT2440 genome (Table S3).

Repeats of overrepresented oligomers can be best seen in the 8–14mers in circle 5 (counting from the inside). These consist of the two oligo classes seen in Fig. 3, the non-coding repeats restricted to narrow genomic areas and the widely distributed oligomers typically found in coding regions. The coding oligomers thus make up a background level of oligomers across the genome (except for in divergent regions). The dark orange clusters of repeat-rich regions consequently consist mainly of non-coding repeats such as the 804

REP elements found in this genome (Aranda-Olmedo *et al.*, 2002). Divergent regions (in blue) are devoid of these common repeats. Instead, they contain local oligomer repeats not found elsewhere in the genome.

Two distinct mismatch regions between tetramers and 8–14mers can be observed in the outermost ring at 0.19–0.22 and 0.92–0.93 Mb. These genomic segments encode large surface adhesion proteins that are made up of numerous long amino acid repeats (Reva and Tümmler, 2008). The regions have a high tetranucleotide oligonucleotide variance, which means many different tetramers are utilized that, together with low occurrence of globally overrepresented 8–14mers, result in the mismatch. The parameter distance is also significantly above average, indicating these regions are markedly different to the genome as a whole on the basis of tetranucleotide patterns.

P. aeruginosa PA14 genome atlas

The *P. aeruginosa* strains are characterized by highly conserved core genomes and a variable set of genomic islands present at a restricted number of insertion points (Klockgether *et al.*, 2007). For example, 95.8% of the PAO1 genome (Stover *et al.*, 2000) is found in strain PA14 (Lee *et al.*, 2006), and the atlas for strain PA14 is thus very similar to the reference strain PAO1 (see Fig. S2).

Pseudomonas aeruginosa PA14 is also highly GC rich (66%), so few regions of above average GC were expected or observed. Repeat regions are more densely clustered in this genome than in *P. putida* KT2440, with some non-divergent regions relatively devoid of repeats. Given the predominance of oligomers spread throughout the genome (the 'background', see Fig. 4) and the fewer obvious repeats, it would appear that some of these widespread oligomers are to a degree clustered in hotspots together with some of the repeats. *Pseudomonas aeruginosa* PA14 has a far greater background coverage of 8–14mers spread throughout the genome, with an average of 19% of each 5 kb region occupied in comparison to 5% in *P. putida*.

It is well known that *P. aeruginosa* strains contain genomic islands with pyocins or phage proteins (Ernst *et al.*, 2003), supporting the divergent regions listed in Table S4. Predicted genomic islands and divergent regions matched well to those already known, including their proximity to well studied recombination and integration sites such as tRNA genes (Klockgether *et al.*, 2007).

The genome atlas is dominated by the 108 kb PAPI-1 pathogenicity island at 5.25–5.36 Mb, which contains many virulence-associated genes (He *et al.*, 2004). This low GC region exhibits a marked difference to the genomic tetranucleotide usage, and a deficiency of globally overrepresented 8–14mers.

Mismatches between 4mers and 8–14mers, which are typical of repeated coding motifs, are also visible. These mismatching regions are, however, much smaller than the large surface adhesion proteins in *P. putida* and are annotated to encode a range of products. For example, the 0.03–0.04 Mb region encodes an ABC transporter, various putative transcriptional regulators and a tryptophan synthase. At 1.71–1.72 Mb several dehydrogenases and six hypothetical proteins are encoded, while the 2.84–2.85 Mb genomic region contains genes for sigma factors, transmembrane sensors, receptors and hypothetical proteins.

Overrepresented peptides

Next, we were curious to discover whether the abundant 8–14mers are just overrepresented oligonucleotides or whether they also encode abundant codons or even frequent peptides. Screening of the list of coding octa- to tetradecanucleotides revealed that each oligomer preferentially starts at one of the three possible positions in the codon (Table 4). Median values were 80%, i.e. 80% of occurrences of one particular oligonucleotide in the genome were in the same coding frame. In other words, the localization of each coding oligomer is strongly skewed to just one of the three possible positions in the reading frame. This bias was the same for all analysed *Pseudomonas* strains and was independent of the total number of overrepresented oligomers in the genome. Hence, the same forces seem to have selected a particular 8–14mer to be overrepresented in the coding regions. The preferential usage of these oligonucleotides is probably driven by global biological and physicochemical constraints that shape structure and function of the chromosome (Baldi and Baisnee, 2000) and, more importantly, by codon usage bias and amino acid frequency at the level of translation (Ikemura, 1985; Sharp and Li, 1987; Chen *et al.*, 2004).

Codon usage could also be influenced by nearest-neighbour codon biases. Although the search of our oligonucleotide data sets did not generally support this hypothesis of a codon context bias, a few overrepresented oligomers were indeed found to be translated in-frame into overrepresented tripeptides (Table S5). Twelve tripeptides were detected from six of the 13 analysed *Pseudomonas* genomes. The 12 peptides were significantly more frequent than predicted from the single amino acid frequencies. Leucine and isoleucine are the most abundant amino acids in the tripeptides (Fig. 5). All peptides carry aliphatic hydrophobic side-chains, four of which are charged (QRL, LID, LIE, IDQ) (Fig. 5). None of the tripeptides were associated with any of the functional categories used in genome annotation. Thus, these short peptides represent global features of the *P. putida*, *P. fluorescens* and *P. syringae* proteomes.

Discussion

To our knowledge, this is the first study that aims to exploit data sets of overrepresented oligonucleotides of intermediate length, i.e. octa- to tetradecanucleotides, for the analysis of the organization of bacterial genomes. Highly overrepresented oligonucleotides were located in high copy numbers in all sequenced genomes of the bacterial genus *Pseudomonas*. This discovery leads us to question why these oligomers are maintained, and why a significant number are surprisingly regularly spaced across most genomic regions.

While to our knowledge no efforts have yet been undertaken to directly test the function of these oligonucleotides of intermediate size, numerous *in silico* analyses have dealt with oligonucleotides of shorter size, i.e. di- to hexanucleotides. Karlin and colleagues (Burge *et al.*, 1992; Karlin *et al.*, 1997; Mrázek *et al.*, 2002) derived statistics for under and over-representation of short oligomers and used these as a basis for postulating dinucleotide signatures for prokaryotic genomes. With the advent of whole bacterial genome sequencing the global analysis of OU in complete sequences of bacterial chromosomes became practically feasible. The patterns of deviations of oligonucleotide frequencies from expectations were shown to be genome signatures, reflecting to some extent the phylogenetic links between microorganisms (Pride *et al.*, 2003; Reva and Tümmler, 2004; Teeling *et al.*, 2004; Bohlin *et al.*, 2008a,b). The usage of oligonucleotides in bacterial sequences was demonstrated to be non-random. Frequencies of the short oligonucleotides depend strongly on their physicochemical properties such as base stacking energy, propeller twist angle, bendability, position preference and protein deformability (Shpigelman *et al.*, 1993; Brukner *et al.*, 1995; Dlakic *et al.*, 2004). Oligonucleotide usage in bacterial genomes is influenced by codon usage biases; however, there may be further yet unknown mechanisms of oligonucleotide selection.

Our previous analysis of sequences of 163 bacterial chromosomes had revealed that most constraints of OU are hidden in di-, tri- and tetranucleotide combinations that vanish with increasing length (Reva and Tümmler, 2005). For example, in the case of a hexamer the four possible heptamers were found to have the same likelihood to occur next in the sequence. Hence, the species-specific OU does not introduce any constraints in the selection of amino acids. Besides some bias in the selection of nearest neighbours, OU does not restrict the protein world in its potential of amino acid sequences. Thus we were taken by surprise to learn within this work that the regional patterns of short oligomer usage in *Pseudomonas* genomes were mirrored by matching patterns of the distributions of the highly overrepresented

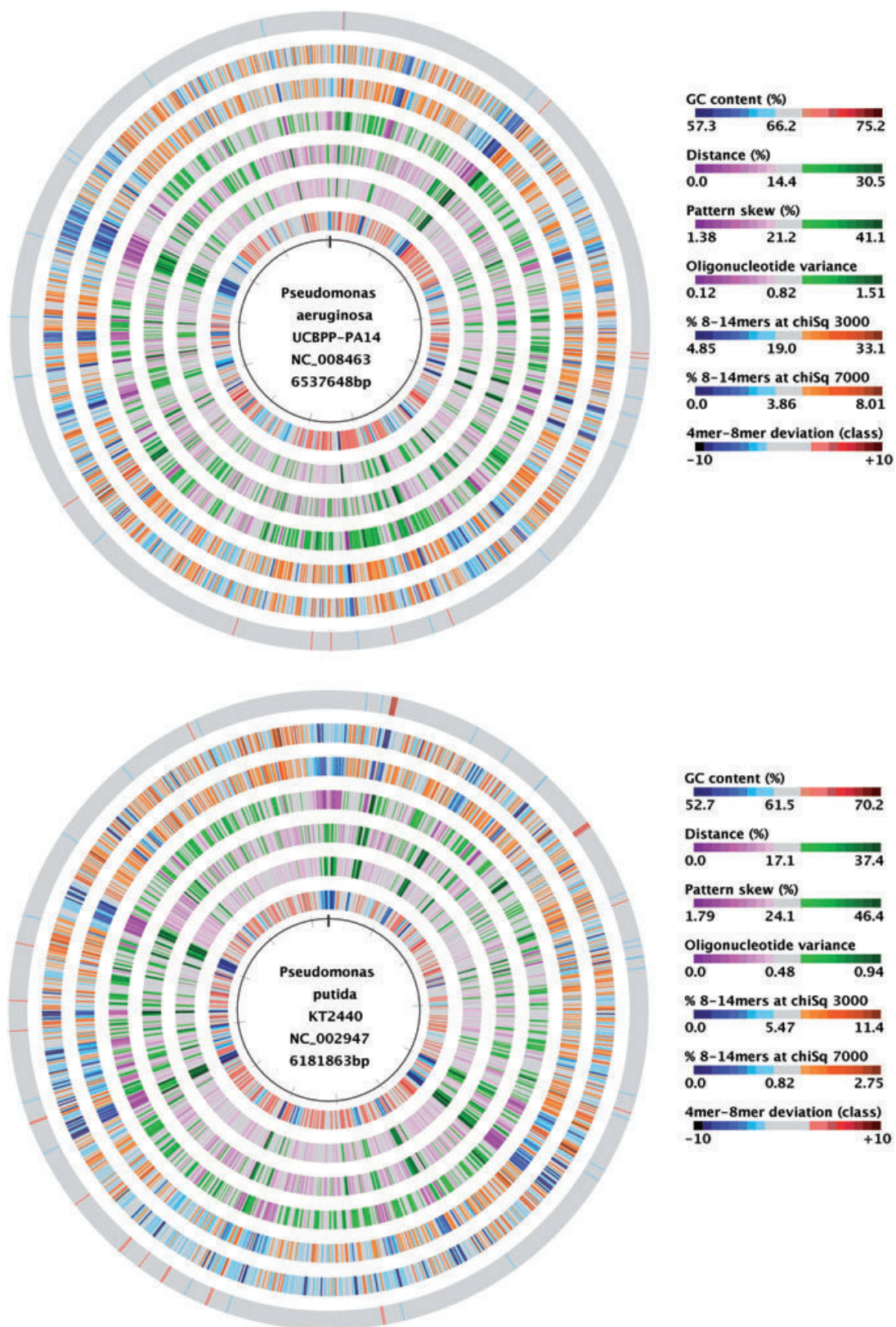


Fig. 4. Genome atlas representations of G+C content, tetranucleotide parameters and overrepresented 8–14mers in *P. aeruginosa* PA14 (upper panel) and *P. putida* KT2440 (lower panel). Increasing divergence from average (up to an extreme value at ± 3 standard deviations) is indicated by progressively darker colours. G+C content and the three tetranucleotide parameters are plotted on the innermost four rings. Distance (second innermost circle) is the distance between global and local sliding window tetranucleotide patterns, Pattern skew (third innermost circle) is the distance between tetranucleotide rankings on direct and reverse strands, and Oligonucleotide variance (fourth innermost circle) is the numerical variance of oligomers, where a lower value indicates tetramer usage is more highly restricted (for example in repeat regions). Rings 5 (χ^2 -threshold 3000) and 6 (χ^2 -threshold 7000) display the number of bases occupied by overrepresented 8–14mers in a certain region, with overlaps only counted once, as a percentage. The outermost ring shows the difference (in classes) between a tetranucleotide parameter, Oligonucleotide variance and the 8–14mers in ring 5. Figures were created with the program JCircleGraph.

octa- to tetradecamers. The statement that OU does not affect coding patterns is globally true, but needs to be modified in case of the most overrepresented words in coding regions. These oligomers are not only abundant because of functional needs to have these peptides frequent in proteins (Table 4, Fig. 5), but they are also abundant because of constraints at the DNA level.

Oligonucleotide usage is highly biased in the extreme subgroup of the most abundant 8–14mers, thus making it a characteristic attribute of the conserved core genome. Correspondingly, the core genome could be differentiated with comparable sensitivity and specificity from the accessory genome by both tetranucleotide usage patterns and the map positions of the longer highly overrepresented 8–14mer markers.

Statistical analysis of oligonucleotide distribution revealed two distinct classes of abundant 8–14mers that virtually do not overlap, i.e. oligomers that cluster in non-coding regions and oligomers that are equally distributed in the coding regions of the core genome (see Table 3). It remains to be seen whether this finding is globally applicable to bacterial genomes or whether it is confined to genomes of the genus *Pseudomonas*.

Each overrepresented coding oligomer is preferentially located in one reading frame (Table 4). With the exception

of a few oligomers that cluster in intragenic amino acid repeats (outermost ring in the genome atlases, Fig. 4), these overrepresented coding oligomers are found throughout the core genome. The preference for one reading frame in genes of the core genome makes this subset of most abundant coding oligomers global features of both genome and proteome.

Most overrepresented non-coding oligomers are part of longer repeats. The 804 copies of the species-specific 35 bp REP element in the *P. putida* genome are a prominent example (Aranda-Olmedo *et al.*, 2002; Weinel *et al.*, 2002a). However, by taking the stringent χ^2 -threshold of 3000 to include an overrepresented oligomer, only the most abundant and conserved repeats were selected. *Pseudomonas aeruginosa* genomes, for example, contain more than 100 copies each of two related 33–35 bp REP elements (Weinel *et al.*, 2002a). As the consensus sequence of the REP is less conserved in *P. aeruginosa* than in *P. putida*, the non-coding 8–14mer constituents of the REPs did not occur frequently enough to show up in Fig. 3. We make this point to clarify that one should not misinterpret the presentations in Fig. 3 as such that non-coding oligomers are diagnostic for *P. putida* and *P. entomophila* and that *P. aeruginosa* is devoid of non-coding oligomer repeats. Figure 3 visualizes the most abundant,

Table 4. Percentage of localization of overrepresented coding 8–14mers in their most common coding frame (start at codon position +1, +2 or +3).

Genome	Number of coding oligomers	Per cent localization of an oligonucleotide in its most common coding frame (%) median (inner quartiles; range)
<i>P. aeruginosa</i> PAO1	459	82.3 (68.6–93.3; 42.3–100.0)
<i>P. aeruginosa</i> PA7	500	82.7 (70.2–92.9; 40.9–100.0)
<i>P. aeruginosa</i> PA14	432	81.8 (68.4–93; 42.1–100.0)
<i>P. entomophila</i> L48	214	83.4 (72.2–91.6; 40.7–99.5)
<i>P. fluorescens</i> PfO-1	91	83.5 (73.3–88; 44–99)
<i>P. fluorescens</i> Pf-5	523	85 (72–93.1; 37.1–100.0)
<i>P. mendocina</i> ymp	385	82.8 (70.2–92.5; 43.2–100.0)
<i>P. putida</i> F1	96	81.1 (73.6–89.3; 46.8–97.6)
<i>P. putida</i> KT2440	87	79.9 (72.4–85.7; 47.8–97.6)
<i>P. stutzeri</i> A1501	131	83.2 (72.6–91.2; 48–100.0)
<i>P. syringae</i> pv. tomato DC3000	30	78.8 (66.2–82.9; 56.5–90.7)
<i>P. syringae</i> pv. syringae B728a	31	79.7 (69.3–84.3; 55–91.1)
<i>P. syringae</i> pv. phaseolicola 1448 A	20	79.3 (74.3–82; 53.8–91.3)

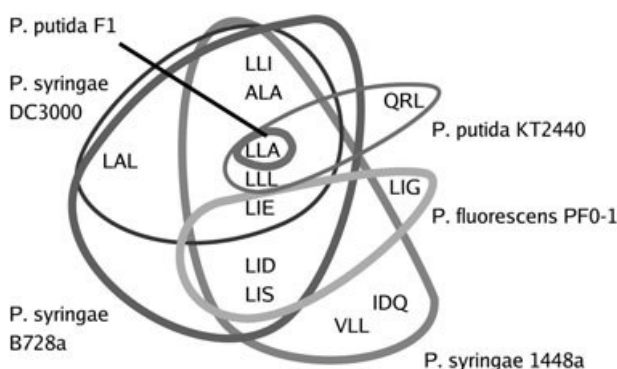


Fig. 5. Statistically significantly overrepresented tripeptides encoded by in-frame overrepresented oligomers from coding regions in six sequenced *Pseudomonas* strains. No overrepresented tripeptides were observed for all *P. aeruginosa* strains, *P. mendocina* ymp, *P. entomophila* L48, *P. fluorescens* Pf-5 or *P. stutzeri* A1501. The repeated presence of leucine (L), the most common amino acid, is marked despite the normalization for expected values.

but not necessarily all common oligomer constituents in repeats.

The biological meaning of the oligomers in our data set is not limited to the highly selected REP elements, which play a role as transposition hotspots (Tobes and Pareja, 2006). One example is the presence of chi-sites in our data set. Chi-sites (5'-GCTGGTGG-3') are involved in the repair of double stranded breaks in chromosomal DNA, modulating the activity of the RecBCD helicase and an exonuclease (Sourice *et al.*, 1998). These sequences were located in four strains in our highly overrepresented χ^2 3000 data set, namely *P. aeruginosa* PAO1 and PA7, *P. fluorescens* Pf-5 and *P. entomophila* L48. Presence of chi-sites was subsequently verified for all strains at the lower χ^2 overrepresentation level of 500, with copy numbers varying from 530 to 1362 (data not shown).

Highly overrepresented 8–14mers that occurred on average once every 10 kb in the genome were classified as markers. All these markers exclusively occurred in the coding sequence, suggesting that their abundance is a consequence of local chromatin constraints, codon usage bias and/or non-random selection of codons for nearest neighbours. The markers' frequency and spatial distribution in the core genome are similar to that of randomly chosen hexa- or heptamers. Hence such markers can be potentially exploited to identify a taxon particularly if the marker is highly overrepresented in the taxon, but rare or even absent in any other taxon.

The emerging wave of high throughput large-scale DNA shotgun sequencing projects (Schuster, 2008) by 454 pyrosequencing (Margulies *et al.*, 2005) or sequencing-by-synthesis technologies (Boyle *et al.*, 2008; Cloonan *et al.*, 2008) will produce massive amounts of primary sequencing data at comparably low cost. Complex biological samples such as bacterial metagenomes thus become amenable to sequencing analysis (Blow, 2008). Sequence reads generated by the new technologies, however, are short. Oligonucleotide markers could gain momentum in identifying the taxa that are present in a bacterial metagenome sample (Venter *et al.*, 2004; Massana *et al.*, 2008; Urich *et al.*, 2008). Knowledge of the frequency and specificity of markers would even allow semi-quantitative estimates about the abundance of at least the major species in the sample. As shown here for taxa of the *Pseudomonas* genus, we have the algorithms at hand to identify, sort and assess specificity of markers. As the next step, marker specificity of a much larger data set of oligomers at less restrictive χ^2 -thresholds will be determined. The set-up of comprehensive marker catalogues and the development and validation of algorithms for data mining will probably be no trivial task. Despite these challenges the approach to identify taxa by oligonucleotide words should be highly worthwhile because the lowest-price-per-nucleotide sequencing technology

yields short reads of just 30–50 bp in length (Hillier *et al.*, 2008). Oligonucleotide word analysis could become the method of choice to quickly resolve the metagenome composition directly from the primary data without any assembly of reads.

Experimental procedures

An in-house command line program, OligoCounter (available at <http://webhost1.mh-hannover.de/davenport/oligocounter/index.html>), was written in Java to count overrepresented oligomers in unannotated fastA files. OligoCounter was used to analyse the whole genomes of 13 *Pseudomonas* sp. genomes available from the NCBI FTP site. These include *P. fluorescens* Pf-5 (GenBank NC_004129, Paulsen *et al.*, 2005), *P. fluorescens* PfO-1 (NC_007492), *P. aeruginosa* PAO1 (NC_002516, Stover *et al.*, 2000), *P. aeruginosa* PA14 (NC_008463, Lee *et al.*, 2006), *P. aeruginosa* PA7 (NC_009656), *P. putida* KT2440 (NC_002947, Nelson *et al.*, 2002), *P. putida* F1 (NC_009512), *P. entomophila* L48 (NC_008027, Vodovar *et al.*, 2006), *P. syringae* DC3000 (NC_004578, Buell *et al.*, 2003), *P. syringae* pv. *syringae* B728a (NC_007005, Feil *et al.*, 2005), *P. syringae* pv. *phaseolicola* 1448 A (NC_005773, Joardar *et al.*, 2005), *P. stutzeri* A1501 (NC_009434, Yan *et al.*, 2008) and *P. mendocina* ymp (NC_009439).

The entire genome is read in before data below threshold values are removed so as to avoid assumptions about the distribution of oligomers throughout the genome. OligoCounter thresholds were set stringently for these large genomes to facilitate graphical overviews of genomic organization granted by the most highly overrepresented oligomers: these had to be present at least 70 times in the genome and be overrepresented at a χ^2 -value of 3000 or more. χ^2 -statistics were calculated according to the formula (Greenwood and Nikulin, 1996):

$$\chi^2 = \frac{(C_{\text{obs}} - C_{\text{exp}})^2}{C_{\text{exp}}}$$

where C_{obs} is the observed count of words and C_{exp} is the expected count of words.

Expected counts of oligomers were derived by a zero order Markov model (Almagor, 1983)

$$E = N \times A^a \times C^c \times G^g \times T^t$$

where N is the genome size in nucleotides, A is the proportion of adenine in the genome and a is the number of adenines in the oligo, and so on for the other bases. The χ^2 -statistic is not intended to be an indicator of statistical significance but merely of level of overrepresentation of each oligomer, otherwise Bonferroni corrections for multiple tests would have been necessary. A random 6 Mb genome with 50% G+C content was generated as a control and demonstrated to have no oligonucleotide biases at a χ^2 -value of 3000.

Results were generated as tab separated plain text files labelled with the species name, GenBank RefSeq identifier, and χ^2 -threshold used. A further Java program, OligoViz, was written to visualize these data in a dual colour scatterplot

format differentiating coding and non-coding areas. Percentage of coding occurrences was also calculated by this program.

In-house scripts were used to analyse the specificity of oligomers. Core oligomer markers were located, which are present on average every 10 000 bp or less in all strains of the respective species. These markers were then screened for in the overrepresented non-*Pseudomonas* chromosomes in the same ' χ^2 3000' data set, of which 407 chromosomes yielded data.

To visualize the interplay on a genomic level between tetranucleotide parameters derived in a 10 kb sliding window and distributions of 8–14mers, the Java program JCircleGraph was developed in-house. G+C content and the three tetranucleotide parameters, namely, Distance between global and local sliding window patterns, Pattern skew as the distance between tetranucleotide rankings on direct and reverse strands, and Oligonucleotide variance (all fully explained in Reva and Tümmler, 2004) were calculated with the Python program OligoWords v1.2 in 10 kb sliding windows with a 5 kb step (Reva and Tümmler, 2005; Ganesan *et al.*, 2008). These parameters are plotted on the inner four rings of the circle. The outer rings contain percent occupancy of each 5 kb genomic region by overrepresented 8–14 bp oligomers at two χ^2 -threshold levels, 3000 and 7000. These statistics are simply the number of bases occupied by 8–14 bp long oligomers, where each base can only be occupied once (i.e. overlaps are ignored), expressed as a percentage. The outermost ring shows the difference (in classes) between a tetranucleotide parameter, Oligonucleotide variance, and the 8–14mers in ring 5, calculated by the following formula:

$$\text{Outermost ring} = (\text{OUVclass} - 8\text{merclass}/2) + 11$$

These parameters mostly negate each other (produce grey) except in 4mer/8mer mismatch regions. JCircleGraph calculates averages and standard deviations for each genome parameter and deviation from average for each of 21 data classes (9 classes up to 3 standard deviations below average, 3 average classes, 9 classes up to 3 standard deviations above average). Each 10 kb data point is then attributed to a class, and plotted in that class' colour. Colours become darker with increasing deviation from average.

To determine if the overrepresented oligomers present in gene-coding regions encode significantly overrepresented peptide motifs, 2 bp from the genomic sequence were added before and after the oligomer. The in-frame (translated) sequence was then converted to amino acids and resultant motifs summed. Expected values for these peptide motifs were calculated using amino acid frequencies taken from the TIGR CMR website (Peterson *et al.*, 2001). Only those peptides overrepresented beyond the $P < 0.05$ significance level are presented (χ^2 -test).

The programs OligoCounter, OligoViz and JCircleGraph are available on the OligoCounter website (<http://webhost1.mh-hannover.de/davenport/oligocounter/index.html>).

Acknowledgements

We are grateful to Jens Klockgether and Frauke Stanke for helpful discussions. The authors would like to thank all

involved in sequencing of the *Pseudomonas* sp. genomes for public access to their data. C.D. is a member of the International Research Training Group 'Pseudomonas Pathogenicity and Biotechnology' (GRK 653/3) sponsored by the Deutsche Forschungsgemeinschaft. O.R. is supported by the National Bioinformatics Network of South Africa.

References

- Allen, T.E., Price, N.D., Joyce, A.R., and Palsson, B.Ø. (2006) Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Comput Biol* **2**: e2.
- Almagor, H. (1983) A Markov analysis of DNA sequences. *J Theor Biol* **104**: 633–645.
- Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L., and Marqués, S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. *Nucleic Acids Res* **30**: 1826–1833.
- Baldi, P., and Baisnee, P.F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* **16**: 865–889.
- Blow, N. (2008) Metagenomics: benchmarks and standards. *Nature* **453**: 687.
- Bohlin, J., Skjerve, E., and Ussery, D.W. (2008a) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* **4**: e1000057.
- Bohlin, J., Skjerve, E., and Ussery, D.W. (2008b) Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* **9**: 104.
- Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Wenig, Z., *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322.
- Brukner, I., Sanchez, R., Suck, D., and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* **14**: 1812–1818.
- Buell, C.R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I.T., Gwinn, M.L., *et al.* (2003) The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. *Proc Natl Acad Sci USA* **100**: 10181–10186.
- Burge, C., Campbell, A.M., and Karlin, S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* **89**: 1358–1362.
- Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA* **101**: 3480–3485.
- Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**: 613–619.
- Diakic, M., Ussery, D., and Brunak, S. (2004) DNA bendability and nucleosome positioning in transcriptional regulation. In *DNA Conformation in Transcription*. Ohyama, T. (ed.) Austin, USA: Landes Bioscience, Springer, pp. 189–203.
- Dos Santos, V.A., Heim, S., Moore, E.R., Strätz, M., and

- Timmis, K.N. (2004) Insights into the genomic basis of niche specificity of *Pseudomonas putida* KT2440. *Environ Microbiol* **6**: 1264–1286.
- Ernst, R.K., D'Argenio, D.A., Ichikawa, J.K., Bangera, M.G., Selgrade, S., Burns, J.L., *et al.* (2003) Genome mosaicism is conserved but not unique in *Pseudomonas aeruginosa* isolates from the airways of young children with cystic fibrosis. *Environ Microbiol* **5**: 1341–1349.
- Feil, H., Feil, W.S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., *et al.* (2005) Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. *Proc Natl Acad Sci USA* **102**: 11064–11069.
- Ganesan, H., Rakitianskaia, A.S., Davenport, C.F., Tümmler, B.T., and Reva, O.N. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**: 333.
- Gelfand, M.S., and Koonin, E.V. (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res* **25**: 2430–2439.
- Greenwood, P.E., and Nikulin, M.S. (1996) *A Guide to Chi-Squared Testing*. New York, USA: Wiley.
- He, J., Baldini, R.L., Déziel, E., Saucier, M., Zhang, Q., Liberati, N.T., *et al.* (2004) The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proc Natl Acad Sci USA* **101**: 2530–2535.
- Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., *et al.* (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* **5**: 183–188.
- Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13–34.
- Joardar, V., Lindeberg, M., Jackson, R.W., Selengut, J., Dodson, R., Brinkac, L.M., *et al.* (2005) Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. *J Bacteriol* **187**: 6488–6498.
- Karlin, S., Mrázek, J., and Campbell, A. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**: 3899–3913.
- 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.
- 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.
- 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.
- Massana, R., Karniol, B., Pommier, T., Bodaker, I., and Béjà, O. (2008) Metagenomic retrieval of a ribosomal DNA repeat array from an uncultured marine alveolate. *Environ Microbiol* **10**: 1335–1343.
- 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.
- Médigue, C., and Moszer, I. (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol* **158**: 724–736.
- Mrázek, J., Gaynon, L.H., and Karlin, S. (2002) Frequent oligonucleotide motifs in genomes of three streptococci. *Nucleic Acids Res* **30**: 4216–4221.
- Nelson, K.E., Weinle, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Martins dos Santos, V.A., *et al.* (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* **4**: 799–808.
- Olsen, G.J., and Woese, C.R. (1993) Ribosomal RNA: a key to phylogeny. *FASEB J* **7**: 113–123.
- Paulsen, I.T., Press, C.M., Ravel, J., Kobayashi, D.Y., Myers, G.S., Mavrodi, D.V., *et al.* (2005) Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat Biotechnol* **23**: 873–878.
- Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H., and Ussery, D.W. (2000) A DNA structural atlas for *Escherichia coli*. *J Mol Biol* **299**: 907–930.
- Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K., and White, O. (2001) The comprehensive microbial resource. *Nucleic Acids Res* **29**: 123–125.
- Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**: 145–155.
- Ramos, J.-L., ed. (2004) *Pseudomonas*, Volumes 1–3. New York, USA: Kluwer Academic/Plenum Publishers.
- Reva, O., and Tümmler, B. (2008) Think big – giant genes in bacteria. *Environ Microbiol* **10**: 768–777.
- Reva, O.N., and Tümmler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**: 90.
- Reva, O.N., and Tümmler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**: 251.
- Reva, O.N., Hallin, P.F., Willenbrock, H., Sicheritz-Ponten, T., Tümmler, B., and Ussery, D.W. (2008) Global features of the *Alcanivorax borkumensis* SK2 genome. *Environ Microbiol* **10**: 614–625.
- Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nat Methods* **5**: 16–18.
- Sharp, P.M., and Li, W.H. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**: 222–230.
- Shpigelman, E.S., Trifonov, E.N., and Bolshoy, A. (1993) CURVATURE: software for the analysis of curved DNA. *Comput Appl Biosci* **9**: 435–440.
- Skovgaard, M., Jensen, L.J., Friis, C., Staerfeldt, H.H., Worning, P., Brunak, S., and Ussery, D.W. (2002) The atlas visualisation of genome-wide information. In *Methods in Microbiology*. Wren, B., and Dorrell, N. (eds). London, UK: Academic Press, pp. 49–63.
- Sourice, S., Biaudet, V., Karoui, M.E., Ehrlich, S.D., and Gruss, A. (1998) Identification of the Chi site of *Haemophi-*

- lus influenzae* as several sequences related to the *Escherichia coli* Chi site. *Mol Microbiol* **27**: 1021–1029.
- Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warren, P., Hickey, M.J., *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**: 959–964.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**: 163.
- Tobes, R., and Pareja, E. (2006) Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics* **7**: 62.
- Urich, T., Lanzén, A., Qi, J., Huson, D.H., Schleper, C., and Schuster, S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* **3**: e2527.
- Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.
- Vodovar, N., Vallenet, D., Cruveiller, S., Rouy, Z., Barbe, V., Acosta, C., *et al.* (2006) Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium *Pseudomonas entomophila*. *Nat Biotechnol* **24**: 673–679.
- Weinel, C., Ussery, D.W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C., and Tümmler, B. (2002a) Comparative genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: orthologs, codon usage, repetitive extragenic palindromic elements, and oligonucleotide signatures. *Genome Lett* **1**: 175–187.
- Weinel, C., Nelson, K.E., and Tümmler, B. (2002b) Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ Microbiol* **4**: 809–818.
- Willenbrock, H., Friis, C., Juncker, A.S., and Ussery, D.W. (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol* **7**: R114.
- Yan, Y., Yang, J., Dou, Y., Chen, M., Ping, S., Peng, J., *et al.* (2008) Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated *Pseudomonas stutzeri* A1501. *Proc Natl Acad Sci USA* **105**: 7564–7569.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. As Fig. 3, with images for every sequenced strain in the genus *Pseudomonas* and a selection from the genus *Bacillus* for comparison.

Fig. S2. As Fig. 4, but with atlases for every sequenced strain in the genus *Pseudomonas* and a selection from the genus *Bacillus* for comparison.

Table S1. Nucleotide sequences of the specific species level markers from Table 2.

Table S2. Divergent regions in three *Pseudomonas* strains as inferred by the lack of overrepresented 8–14 bp oligomers in Fig. 3. Genome position is presented in megabases.

Table S3. Genomic regions with GC diverging by more than two standard deviations from the mean in *P. putida* KT2440. '+', '0' or '-' indicate an above average, average, or below average value respectively.

Table S4. Genomic regions with GC diverging by more than two standard deviations from the mean in *P. aeruginosa* strain PA14. '+', '0' or '-' indicate an above average, average, or below average value respectively.

Table S5. Statistically significantly overrepresented tripeptides (χ^2 , $P < 0.05$) encoded by overrepresented oligomers in the genus *Pseudomonas*. Observed values are presented, with expected values (see *Experimental procedures*) following in brackets.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Comparative Genomics of Green Sulfur Bacteria

Colin Davenport¹, David W. Ussery², and Burkhard Tümmeler¹

¹Klinische Forschergruppe, Klinik für Pädiatrische Pneumologie und Neonatologie, OE6711,
Medizinische Hochschule Hannover, Carl-Neuberg-Strasse 1, D-30625 Hannover, Germany

²Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

Correspondence: Burkhard Tümmeler, Klinische Forschergruppe, OE 6710, Medizinische Hochschule
Hannover, Carl-Neuberg. Str. 1, D-30625 Hannover , Germany, phone: +49-511-5322920, FAX:
+49-511-5326723, email: tuemmler.burkhard@mh-hannover.de

Abstract

Eleven completely sequenced *Chlorobi* genomes were compared in oligonucleotide usage, gene contents and synteny. The green sulfur bacteria are equipped with a core genome that sustains their anoxygenic phototrophic lifestyle by photosynthesis, sulfur oxidation and CO₂ fixation. Whole-genome gene family and single gene sequence comparisons yielded similar phylogenetic trees of the sequenced chromosomes indicating a concerted vertical evolution of large gene sets. Chromosomal synteny of genes is not preserved in the phylum *Chlorobi*. The accessory genome is characterized by anomalous oligonucleotide usage and endows the strains with individual features for transport, secretion, cell wall, extracellular constituents and a few elements of the biosynthetic apparatus. Giant genes are a peculiar feature of the genera *Chlorobium* and *Prosthecochloris*. The predicted proteins have a huge molecular weight of 10⁶ and are probably instrumental for the bacteria to generate their own intimate (micro)environment.

Keyword: comparative genomics, chlorobaculum, chlorobium, genome atlas, oligonucleotide usage

List of abbreviations:

Bchl, bacteriochlorophyll

Chl, chlorophyll

CRISPRs, clustered regularly interspaced short palindromic repeats

FMO protein, Fenna-Matthews-Olson protein

GRV, global relative variance

GSB, green sulfur bacteria

LCB, locally collinear block

OU, oligonucleotide usage

RV, relative variance less than 10

TCA, tricarboxylic acid

Introduction

The green sulfur bacteria (GSB; phylum *Chlorobi*, family *Chlorobiaceae*) are obligate anoxygenic phototrophs that occur in high densities in a variety of anaerobic low-light environments (Castenholz et al. 1990; Wahlund et al., 1991) containing sources of light and reductant, usually reduced sulfur compounds (Chan et al. 2008). Based on the criteria of GC content (% GC), 16S rDNA and *fmo* (Fenna-Matthews-Olson protein) gene sequences, the family *Chlorobiaceae* has been divided into the genera *Chlorobium*, *Prosthecochloris* and *Chlorobaculum* (Alexander et al. 2002; Imhoff 2003). Typical habitats of GSB are hot springs, hydrothermal vents, sediments, hypolimnia, and microbial mats (Overmann and Garcia-Pichel 2006).

GSB have characteristic light-harvesting structures, the chlorosomes, which are attached to the cytoplasmic membrane carrying the photosynthetic reaction center which contain chlorophylls as essential cofactors (Bryant and Frigaard 2006; Chew and Bryant 2007; Tavano and Donohue 2006). *Chlorobi* synthesize Chl a and BChl a and one of three so-called *Chlorobium* chlorophylls, which are denoted BChl c, d, and e. GSB perform autotrophic CO₂ fixation via the reductive tricarboxylic acid (TCA) cycle by oxidizing various inorganic sulfur compounds as electron donors. Thereby they utilize various combinations of sulfide, elemental sulfur, and thiosulfate and sometimes also ferrous iron and hydrogen. Frigaard and Dahl (2009) recently reviewed the dissimilatory and assimilatory metabolism of inorganic sulfur compounds in GSB and presented a genome-based survey of the distribution and phylogenies of genes involved in oxidation of sulfur compounds in completely sequenced strains.

The first completely sequenced *Chlorobi* genome was that of *Chlorobaculum* (*Chlorobium*) *tepidum* TLS, an isolate from high-sulfide hot springs in New Zealand (Wahlund et al. 1991). The annotation uncovered the bioenergetic capacity of the organism, i.e. the components of light harvesting, electron transfer, and sulfur oxidation, the biosynthesis of chlorophyll, bacteriochlorophyll, carotenoids, isoprenoid quinones and the proteins involved in CO₂ fixation (Eisen et al. 2002; Frigaard et al. 2003). The *in silico* analysis revealed further remarkable features such as the multiple duplication of genes for photosynthesis and sulfur metabolism, the close similarity of many genes to those of archaea, the low number of regulatory genes, the limited capabilities for the assimilation of organic compounds and an incomplete set of tRNA synthetases (Eisen et al. 2002).

Since the initial report in 2002 ten further complete genome sequences of *Chlorobi* strains have been deposited in GenBank. With this increasing number of complete genome sequences at hand, it makes sense to complement genome annotation with the analysis of genome organization and structure. Here we report on the *in silico* comparison of the global features of 11 completely sequenced *Chlorobi* genomes.

Materials and Methods

Genome sequences

Genome sequences and genome statistics were obtained from the NCBI Genome database (<http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genome>).

Genome Atlases

Oligonucleotide usage. To visualise *Chlorobi* genomes using oligonucleotide parameters the Java program JCircleGraph from the OligoCounter package (Davenport et al. 2009) was used. GC content and the three tetranucleotide parameters, namely, Distance between global and local sliding window patterns, Pattern skew as the distance between tetranucleotide rankings on direct and reverse strands, and Oligonucleotide variance (all fully explained in Reva and Tümmler 2004) were calculated with the Python program OligoWords v1.2 in 10 kb sliding windows with a 5 kb step (Reva and Tümmler 2005; Ganesan et al. 2008). These parameters are plotted on the inner four rings of the circle. The outer rings contain percent occupancy of each 5 kb genomic region by overrepresented 8-14 bp oligomers at two χ^2 - threshold levels, 500 and 1,200. These statistics are simply the number of bases occupied by 8-14 bp long oligomers, where each base can only be occupied once (i.e. overlaps are ignored), expressed as a percentage. The outermost ring shows the difference (in classes) between a tetranucleotide parameter, Oligonucleotide variance, and the 8-14mers in ring 5, calculated by the following formula:

$$\text{Outermost ring} = 0.5 (\text{OUVclass} - \text{8merclass}) + 11$$

These parameters mostly negate each other (produce grey) except in 4mer / 8mer mismatch regions. JCircleGraph calculates averages and standard deviations for each genome parameter and deviation from average for each of 21 data classes (9 classes up to 3 standard deviations below

average, 3 average classes, 9 classes up to 3 standard deviations above average). Each 10 kb data point is then attributed to a class, and plotted in that class' colour. Colours become darker with increasing deviation from average.

Divergent regions were selected using the Filter function in the SeqWord Genome Browser (Ganesan et al 2008). The following parameters were used, as recommended on the SeqWord website; relative variance (RV) less than 10, global relative variance (GRV) of more than 15, and distance (D) of more than 29. Briefly, RV is the tetranucleotide pattern calculated normalized by frequencies of mononucleotides in the analyzed genomic fragment, and GRV is the variance of a tetranucleotide pattern of the same fragment normalized by frequencies of mononucleotides in a complete genome. In other words, the difference between RV and GRV is local versus global pattern normalisation.

BLAST atlas. The BLAST atlas (Hallin et al. 2008) used all the provided annotated coding sequences (or proteins) of the reference genome *C. tepidum* TLS, and compared each of those with ten *Chlorobi* genomes. The BLASTatlas method uses the software BLASTALL v. 2.2.11 for the search, and in BLAST terminology, the reference genome constitutes the query whereas each other genome in the comparison corresponds to the database. BLAST searches with a non-stringent E-value cut-off of 0.01 were used to identify the best alignments between the reference sequence protein and the *Chlorobi* genome in question. Each protein in the reference genome was compared to all the proteins in the query set—regardless of orientation or location. Once identified, the single best pairwise alignment for each of the reference sequences was obtained and included in the map. DNA structural properties (Pedersen et al. 2000), repeats (Skovgaard et al. 2002) and measures of base composition (Jensen et al. 1999) of the reference genome were incorporated into the inner circles of the atlas. Once the BLAST searches had been carried out and remapped to the reference genome and custom data and DNA properties had been collected, an XML configuration file was composed which contained all these data and the layout of the atlas. This file was then sent to the GeneWiz software (Pedersen et al. 2000) which produced a PostScript document.

Mauve whole genome alignments.

The Mauve whole genome aligner v 2.2.0 (Darling et al. 2004) was used to globally align and find locally collinear blocks conserved in the genomes of (i) *Chlorobaculum tepidum* and

Chlorobaculum parvum, (ii) extending group i to include two *Chlorobium* species, *Chlorobium luteolum* and *Chlorobium chlorochromatii*, and (iii) all five *Chlorobium* species. Groups were selected based on rDNA similarity and taxonomy by Imhoff (2003). Default alignment settings were used: alignment by Muscle 3.6, minimum island size 50, default seed weight 15.

Repeats and 16S rRNA

CRISPR statistics were taken from the CRISPRdb database (Grissa et al. 2007), while counts of 16S rRNA were retrieved from the GenomeAtlas website (Hallin and Ussery 2004).

Phylogenetic reconstructions

Single proteins. Amino acid sequences were retrieved from the NCBI website and aligned with the multiple sequence alignment program MAFFT in JalView (Waterhouse et al. 2009). Alignments were trimmed and trees built using the Neighbour Joining algorithm in ClustalX with 100 bootstrap resamplings (Larkin et al. 2007). Trees were then visualised using Dendroscope (Huson et al 2007).

The following proteins were used to construct phylogenetic trees:

Bacteriochlorophyll A protein (FMO), accession numbers NP_662384, YP_001999184, CAC26966, Q46135, YP_912328, YP_375402, YP_001943725, ZP_01386023, YP_002017957, YP_001130839, YP_379632, YP_002016260, YP_001995626.

Protochlorophyllide reductase iron-sulfur ATP-binding proteins, accession numbers NP_663024, YP_001997848, YP_380107, YP_374148, ZP_01386402, YP_001129807, YP_912798, YP_001944195, YP_002017276, YP_002016684, YP_001960643, YP_001996281.

Whole-genome gene family clustering. The set of gene families for the collection of *Chlorobium* genomes were found by first aligning all proteins in each genome against all proteins in the query genome with BLAST. A hit was considered significant only if the alignment covered at least 50% of the lengths of the query sequence, and contained at least 50% identities over the query sequence (Tettelin et al. 2005). A pair of genes having significant reciprocal best hits belong to the same gene family. A pan-matrix was constructed, containing one row for each genome and one column for each gene family. Cell (i,j) in this matrix is 1 if genome i has a member in gene family j, 0 otherwise. The hierarchical clustering was based on the Manhattan distance between genomes using

this matrix, and using the complete linkage function (Snipen et al. 2009). Shorter distances represent genomes with many gene families in common, and larger distances reflect genomes with fewer gene families in common. Confidence of the tree topology was assessed using 100 bootstraps.

Results and Discussion

Global genome features

The genomes of the 11 completely sequenced GSB are each composed of a single, circular molecule of 1,966,858 bp (*Prosthecochloris vibrioformis* DSM265) to 3,293,456 bp (*Chloroherpeton thalassium* ATCC35110) (Table 1). The median genome size of *Chlorobi* of about 2.5 Mbp is considerably lower than that of free-living environmental proteobacteria or firmicutes which correlates with the small number of genes for regulatory responses in the former and the higher number in the latter. The larger the genome, the more regulators per gene are encoded (Cases et al., 2003).

The 11 genomes have a G+ C content between 44% and 56% (median: 50%) (Table 1). 82% to 91% of sequence (median: 87%) are coding (Table 1). These values are close to the average among all bacteria.

Repeats

Clustered regularly interspaced short palindromic repeats (CRISPRs) constitute a particular family of tandem repeats found in a wide range of prokaryotic genomes (half of eubacteria and almost all archaea). They consist of a succession of highly conserved regions varying in size from 23 to 47 bp, separated by similarly sized unique sequences (spacer) of usually viral origin (Grissa et al. 2007). CRISPRs have been shown to play a role in acquired prokaryotic resistance to viruses, perhaps via an RNAi like pathway (Jansen et al 2002, Sorek et al 2008). According to the CRISPRs database (Grissa et al. 2007) all fully sequenced species contain CRISPRs, with at most five copies identified in *Chlorobium phaeobacteroides* DSM266 and *Prosthecochloris aestuarii* DSM271 (Table 1). While copy numbers of CRISPRs may appear low, they are actually well within the normal range. The archaeon *Methanocaldococcus jannaschii* DSM2661 contains 20 CRISPRs, the largest number known to date, which is distinctly more than the normal range for bacteria of 0-8 (Grissa et al. 2007).

Pseudogenes

The accumulation of pseudogenes and large non-coding genomic regions is generally taken as evidence for genome decay. Pseudogene counts vary greatly from 0 in *Chlorobium chlorochromatii* CaD3 to 212 in *Chlorobium phaeoclathratiforme* BU-1 (Table 1). However, there is little reason to suggest that genome decay is operating in the *Chlorobi* genera, because 82% to 91% of sequence (median: 87%) are coding (Table 1). These values are close to the average among all bacteria. In contrast, *Mycobacterium leprae* TN, the paradigm of genome decay, is of similar genome size to the *Chlorobi* at 3 Mbp but contains 1116 pseudogenes and only 49 % of the genome contains protein coding genes (Bentley and Parkhill 2004). It should also be noted that different annotation teams place widely varying emphasis on detecting pseudogenes.

Oligonucleotide usage

The compositional biases of short oligonucleotides are known to represent genomic signatures and to contain a phylogenetic signal (Pride et al. 2003; Reva and Tümmler 2004; Teeling et al. 2004; Bohlin et al. 2008a; 2008b; Reva et al. 2008). Frequencies of mono- to hexanucleotides depend on their physicochemical properties such as base-stacking energies, propeller twist or bendability (Baldi and Baisnee 2000), codon usage bias (Sharp and Li 1987; Willenbrock et al. 2006) and biological constraints of e.g. DNA replication and repair, transcriptional or translational control, or restriction-modification systems (Gelfand and Koonin, 1997; Karlin et al., 1997). Moreover, datasets of the extreme subgroup of the most abundant oligonucleotides of intermediate length, i.e. octa- to tetradecanucleotides, have been qualified as a characteristic attribute of the conserved core genome (Davenport et al. 2009).

Thus we constructed Genome Atlases of tetranucleotide usage and of abundance of overrepresented 8mers- 14mers for all sequenced *Chlorobi* genomes (Figure 1, Online Resource). The innermost circles visualize the variation of the GC content along the chromosome in 10 kb sliding windows. *Chlorobaculum (Chlorobium) tepidum* TLS harbours GC – rich regions around the origin of replication, the segments around 0.42 Mbp, 0.88 Mbp and in a large area from 1.5 to 1.8 Mbp. In these segments of 60% GC or more the local tetranucleotide usage is very similar to the global tetranucleotide usage (distance, second innermost circle) and the individual 256 tetranucleotides occur

at similar frequencies on both strands (pattern skew, third innermost circle). In contrast, regions with about 50% GC (marked blue in the innermost circle) are characterized by high local distance and pattern skew (marked green in the second and third innermost circles). Thus, low GC content (compared to the genomic average) is the major determinant for atypical oligonucleotide usage in *C. tepidum*. A different behaviour was seen in the *Chlorobium chlorochromatii* CaD3 genome (Figure 1, lower panel). Both the regions with low and high GC content deviate in their tetranucleotide usage from chromosomal average and are characterized by high distance, high pattern skew and in most segments also by high oligonucleotide variance. The most anomalous segments with the highest GC content encode three giant genes (coordinates 717094-827511, 829510-891450, 1656265-1705200; Table 2) the first one being the largest bacterial gene known to date (Reva and Tümmler 2008).

Giant genes

Genomes of the genera *Chlorobium* and *Prosthecochloris*, but none of the yet sequenced genomes of the genus *Chlorobaculum* encode giant genes of more than 15 kb in size (Table 2). All but the two largest genes were annotated to encode VCBS outer membrane adhesins. Although no wet lab data is available, the evidence from the *in silico* sequence comparisons is strong that the genetic repertoire of the genera *Chlorobium* and *Prosthecochloris* includes genes that encode giant adhesins. The predicted proteins have a huge molecular weight of 10^6 and are probably instrumental for the bacteria to generate their own intimate (micro)environment (Reva and Tümmler 2008).

Genomic islands and atypical regions

Genomic segments with atypical oligonucleotide usage are indicative of repeats, non-coding regions, large genes, ancient conserved genes, mobile genetic elements or genomic islands, the latter two of which are acquired by horizontal gene transfer (Reva and Tümmler 2005). Table 3 lists the regions of *Chlorobi* genomes which show divergent oligonucleotide parameters. The functional categories encoded by the segments with atypical oligonucleotide usage are similar across most GSB genomes. The islands encode elements of genetic mobility (integrase, transposases, helicases) and restriction systems, transporters, secretion systems, cellular appendages, operons for the biosynthesis of polysaccharides and a few enzymes (Table 3). Thus the GSB have putatively acquired auxiliary

functions and some genes associated with their typical lifestyle as anoxygenic phototrophs (Raymond et al. 2002) by horizontal transfer.

Low synteny of GSB genomes

Regions of genomic similarity called locally collinear blocks of at least 50 bp (LCBs) were examined in the *Chlorobi* using the genome alignment procedure Mauve (Darling et al. 2004). Each locally collinear block of at least 50 bp in size is a homologous region of sequence shared by two or more of the genomes being aligned, and does not display rearrangements of homologous sequence.

Our first observation was that few blocks are aligned when comparing representatives from the *Chlorobium* and *Chlorobaculum* genera (Figure 2A), or even within one genus such as *Chlorobium* (Figure 2B). This implies most LCBs are not orthologous across all five genomes. Unaligned regions may be lineage-specific sequence, be paralogous repetitive regions or may have been rearranged. Furthermore novel sequence introduced to a subset of genomes will not be aligned as they do not have phylogenetic anchors across all sequences (Darling et al. 2004).

The largest block conserved across *Chlorobaculum* and *Chlorobium* is between ca. 650 and 700 kbp in *C. tepidum* (Figure 2a). This block contains conserved stretches of housekeeping genes, for example a NUDIX hydrolase, cysteine synthases, biosynthesis proteins, and ABC and sulfate transporters.

If just two closely related species are examined, such as *C. tepidum* and *C. parvum*, it is notable that almost all blocks are orthologous and conserved (Figure 2C). This confirms the highly relatedness of these two strains based on molecular evidence and the repositioning of *Chlorobium tepidum* as *Chlorobaculum tepidum* (Imhoff 2003). In other words, only a few regions in the *C. tepidum* genome (topmost line of blocks) are lineage specific and have no counterparts in *C. parvum*. For example, a block at ca. 1.63 Mb to 1.68 Mb is not present in *C. parvum*. However, quite a few blocks are inverted or resized. While syntenic block order has been generally conserved the entire genomes have been inverted in respect to one another, i.e. the first blocks from *C. tepidum* are almost always located at the end of the *C. parvum* genome.

BLAST Atlas: conservation of gene content among Chlorobi

Synteny between *Chlorobi* genomes was low. This finding was not unexpected because comparative genome analyses in other branches of the Bacteria yielded similar results that global chromosomal organization is conserved, but that synteny is typically observed only within species or among closely related taxa, but not across the whole phylum (Rocha 2008). If the arrangement of genes has been disrupted during speciation and natural history, gene content itself, however, may still have been preserved. Hence, taking the initially sequenced *C. tepidum* TLS genome as reference, we asked to what extent the coding repertoire is conserved among *Chlorobi*.

The BLAST atlas (Figure 3) used all the provided annotated coding sequences of the reference genome *C. tepidum* TLS, compared each of those with the ten other *Chlorobi* genomes and mapped the best alignment between the reference sequence protein and the *Chlorobi* genome in question onto the atlas. The BLAST atlas visualizes a similar degree of conservation of homologues among the sequenced *Chlorobi*. With the exception of the phylogenetically more distant *Chloroherpeton thalassium* (circle in magenta) that carries less homologues in its genome, the set of orthologues is roughly conserved among all *Chlorobi* (Figure 3). The gaps with no or few orthologues to *C. tepidum* ORFs often coincide with the map position of repeats (third and fourth innermost circles), segments with low GC content (innermost circle) and/or areas with extreme values of base stacking energy or position preference (sixth and seventh innermost circle). The two latter parameters are scales for the dinucleotide interaction of bases and local flexibility of DNA and depend on the species-specific compositional biases of di- to hexanucleotide selection (Pedersen et al. 2000; Reva and Tümmeler 2004). Hence it is interesting to note that the segments with atypical local DNA structure in the *C. tepidum* chromosome match with the areas of no or few homologues in other *Chlorobi*. In other words, the anomalous chromosomal regions with atypical DNA structure and oligonucleotide usage contain the *C. tepidum*-specific accessory genes which were most likely acquired by horizontal gene transfer. On the contrary, the large chromosomal areas with *C. tepidum* – typical DNA structure and oligonucleotide usage -represent the core genes that encode the basic genetic repertoire homologues of which are found in all GSB. The core genome contains, amongst others, the genes for the biosynthesis

of the components of the photosynthetic apparatus, sulfur oxidation and carbon assimilation that are key for the lifestyle of green sulfur bacteria (Eisen et al. 2002; Frigaard et al. 2003).

Phylogenetic trees

Next, a dendrogram of the relatedness of sequenced GSB strains was constructed on the basis of whole-genome comparisons of all proteins (Figure 4). A summary figure of this data showing pairwise comparisons of proportions of proteins in common to both strains is provided in the Online Resource. As expected, the genome of the GSB *C. thalassium* (Gibson et al 1984) was found to be most distant from genomes belonging to the genera *Chlorobium*, *Chlorobaculum* and *Prosthecochloris*. The figure confirms the classification of Imhoff (2003) based on 16S rDNA and *fmo* sequences that (a) strains of the genus *Pelodictyon* were transferred to the genus *Chlorobium*, that (b) *C. limicola* and *C. phaeobacteroides* belong to the same group (Alexander et al. 2002) and that (c) *C. tepidum* and *C. parvum* were correctly assigned to a new genus named *Chlorobaculum*. In contrast, the segregation of *C. phaeobacteroides* BS1 with *Prosthecochloris aestuarii* DSM271, the type species of the genus *Prosthecochloris*, was unexpected. The BS1 bacterium from the Black Sea chemocline is extraordinarily low-light adapted and has not to our knowledge been discovered in any other ecosystem. *P. aestuarii* DSM271 was originally isolated from the hydrogen sulfide-rich mud of a shallow lagoon with elevated salt concentration (Gorlenko 1970), whereas *C. phaeobacteroides* DSM266 was isolated from the anoxic sulfide containing water 19.5 m below surface of meromictic Lake Blankvann in Norway (Pfennig 1968). Thus all three strains were thriving in GSB-typical habitats, but the BS1 strain was derived from the most extreme environment. The whole-genome BLAST comparison now strongly suggests that the coding capacity of BS1 is more related to DSM271 than to DSM266 and hence its classification as a *C. phaeobacteroides* should be reconsidered.

The molecular taxonomy of *Chlorobi* has so far been based on single gene comparisons (Alexander et al. 2002, Imhoff 2003), and we reproduced this procedure by phylogenetic tree constructions of two proteins, the FMO protein and the protochlorophyllide reductase iron sulfur ATP-binding protein chlL (Bryant and Frigaard 2006; Tavano and Donohue 2006) for the 11 sequenced genomes. As shown in Figure 5, *C. thalassium* presents the most highly diverged sequences. A *Chlorobaculum* cluster is near the *C. thalassium* branch, and a cluster of marginally resolved

Chlorobium species are found at the bottom of the panels. *Prosthecochloris vibrioformis* DSM265 is found within the *Chlorobium* cluster in accordance with the whole-genome comparison and Imhoff's decision (2003) to rename DSM265 as *Chlorobium phaeovibrioides*. The BS1 strain again clusters with *P. aestuarii* DSM271 which supports the conclusion from whole-genome gene family clustering (Figure 4) that the monophyletic BS1 strain with its unique 16S rDNA sequence belongs to the genus *Prosthecochloris* rather than to *Chlorobium*. In summary, tree reconstruction from the two full length gene sequences agrees remarkably well with the whole-genome comparisons, although the discriminatory power to reliably differentiate *Chlorobium* species is naturally lower.

Conclusions

The taxa of the phylum *Chlorobi* share a large set of orthologues to accomplish the demanding task of photosynthesis, sulfur oxidation and CO₂ fixation. Whole-genome gene family as well as *chlL* and *fmo* gene sequence comparisons yielded similar phylogenetic trees of the sequenced chromosomes. This finding is consistent with a concerted vertical evolution of large gene sets. Chromosomal synteny of genes, however, was not preserved in the phylum which may partly attributed to the large extent of gene duplication of genes involved in sulfur metabolism and the biosynthetic pathways for photosynthesis (Eisen et al. 2002). The core genome is interrupted by small segments of the flexible strain-specific accessory genes that are characterized by anomalous local DNA structure and oligonucleotide usage. Mostly acquired by horizontal gene transfer, the accessory genes endow the strains with individual features for biosynthesis, transport, secretion, cell wall and extracellular constituents. In summary, green sulfur bacteria are equipped with a core genome that, among others, sustains their anoxygenic phototrophic lifestyle and a small accessory genome that mainly provides auxiliary functions for communication with the outside world.

Acknowledgments

A stipend provided by the Deutsche Forschungsgemeinschaft to C.D. (GRK653/3) is gratefully acknowledged. Part of this work was supported by a grant from the Danish Center for Scientific Computing.

References

- Alexander B, Andersen JH, Cox RP, Imhoff JF (2002) Phylogeny of green sulfur bacteria on the basis of gene sequences of 16S rRNA and of the Fenna-Matthews-Olson protein. *Arch Microbiol* 178:131-40.
- Baldi P, Baisnee PF (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* 16: 865-889.
- Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. *Annu Rev Genet* 38:771-92.
- Bohlin J, Skjerve E, Ussery DW (2008a) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* 4: e1000057.
- Bohlin J, Skjerve E, Ussery DW (2008b) Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* 9: 104.
- Bryant DA, Frigaard NU (2006) Prokaryotic photosynthesis and phototrophy illuminated. *Trends Microbiol* 14:488-96.
- Cases I, de Lorenzo V, Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* 11:248-53.
- Castenholz R W, Bauld J Jorgenson BB (1990) Anoxygenic microbial mats of hot springs: thermophilic *Chlorobium* sp. *FEMS Microbiol Ecol* 74:325–336.
- Chan LK, Weber TS, Morgan-Kiss RM, Hanson TE (2008) A genomic region required for phototrophic thiosulfate oxidation in the green sulfur bacterium *Chlorobium tepidum* (syn. *Chlorobaculum tepidum*). *Microbiology* 154:818–829.
- Chew AG, Bryant DA (2007) Chlorophyll biosynthesis in bacteria: the origins of structural and functional diversity. *Annu Rev Microbiol* 61:113-29.
- Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394-403.
- Davenport CF, Wiehlmann L, Reva ON, Tümmler B (2009) Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ Microbiol* 11:1092-104.

- Eisen JA, Nelson KE, Paulsen IT, Heidelberg JF, Wu M, Dodson RJ, Deboy R, Gwinn ML, Nelson WC, Haft DH, Hickey EK, Peterson JD, Durkin AS, Kolonay JL, Yang F, Holt I, Umayam LA, Mason T, Brenner M, Shea TP, Parksey D, Nierman WC, Feldblyum TV, Hansen CL, Craven MB, Radune D, Vamathevan J, Khouri H, White O, Gruber TM, Ketchum KA, Venter JC, Tettelin H, Bryant DA, Fraser CM (2002) The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proc Natl Acad Sci USA* 99:9509–9514.
- Frigaard NU, Chew AG, Li H, Maresca JA, Bryant DA (2003) *Chlorobium tepidum*: insights into the structure, physiology, and metabolism of a green sulfur bacterium derived from the complete genome sequence. *Photosynth Res* 78:93-117.
- Frigaard NU, Dahl C (2009) Sulfur metabolism in phototrophic sulfur bacteria. *Adv Microb Physiol* 54:103-200.
- Ganesan H, Rakitianskaia AS, Davenport CF, Tümmeler B, Reva ON (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* 9:333.
- Gelfand MS, Koonin EV (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res* 25: 2430-9.
- Gibson J, Pfennig N, Waterbury JB (1984) *Chloroherpeton thalassium* gen. nov. Et spec. nov., a non-filamentous, flexing and gliding green sulfur bacterium. *Arch Microbiol* 138:96-101.
- Gorlenko VM (1970) A new phototrophic green sulfur bacterium-- *Prosthecochloris aestuarii* nov. gen. nov. sp. *Z Allg Mikrobiol* 10:147-9.
- Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35(Web Server issue):W52-7.
- Hallin PF, Ussery DW (2004) CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* 20:3682-3686.
- Hallin PF, Binnewies TT, Ussery DW (2008) The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology. *Mol Biosyst* 4:363-71.

- Huson DH, Richter DC, Rausch C, DeZulian T, Franz M, Rupp R (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8:460.
- Imhoff JF (2003) Phylogenetic taxonomy of the family Chlorobiaceae on the basis of 16S rRNA and fmo (Fenna-Matthews-Olson protein) gene sequences. *Int J Syst Evol Microbiol* 53:941-51.
- Jansen R, Embden JD, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43:1565-75.
- Jensen, LJ, Friis C, Ussery DW (1999) Three Views of Microbial Genomes. *Res Microbiol* 150:773–777.
- Karlin S, Mrazek J, Campbell A (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* 179: 3899-3913.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-8.
- Overmann J, Garcia-Pichel F (2006) The phototrophic way of life. In: Dworkin M et al (eds) *The prokaryotes*, 3rd edn. New York, Springer, pp 32–85.
- Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW (2000) A DNA structural atlas for *Escherichia coli*. *J Mol Biol* 299:907–930.
- Pfennig N (1968) *Chlorobium phaeobacteroides* nov. spec. and *C. phaeovibrioides* nov. spec., zwei neue Arten der grünen Schwefelbakterien. *Arch Microbiol* 63: 224-226.
- Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 13: 145-155.
- Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science* 298:1616-1620.
- Reva ON, Tümmler B (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* 5:90.
- Reva ON, Tümmler B (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* 6:251.
- Reva O, Tümmler B (2008) Think big - giant genes in bacteria. *Environ Microbiol* 10: 768-77.

- Rocha EP (2008) The organization of the bacterial genome. *Annu Rev Genet* 42:211-33.
- Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 4: 222-30.
- Skovgaard M, Jensen LJ, Friis C, Stærfeldt HH, Worning P, Brunak S, Ussery D (2002) The Atlas Visualisation of Genome-wide Information. *Methods Microbiol* 33:49–63.
- Snipen L, Almøy T, Ussery DW (2009) Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* 10: in press (August 2009).
- Sorek R, Kunin V, Hugenholtz P (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* 6:181-6. Tavano CL, Donohue TJ (2006) Development of the bacterial photosynthetic apparatus. *Curr Opin Microbiol* 9:625-31.
- Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5: 163.
- Tettelin H, Maignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* 102:13950-5.
- Wahlund TM, Woese CR, Castenholz RW, Madigan MT (1991) A thermophilic green sulfur bacterium from New Zealand hot springs, *Chlorobium tepidum* sp. nov. *Arch Microbiol* 156:81–90.
- Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189-91.

Willenbrock H, Friis C, Juncker AS, Ussery DW (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol* 7: R114.

Figure Legends

Fig. 1 Genome atlas representations of GC content, tetranucleotide parameters and overrepresented 8-14mers in *Chlorobaculum tepidum* TLS (upper panel) and *Chlorobium chlorochromatii* CaD3 (lower panel). The coordinates of two adjacent giant genes are annotated in *C. chlorochromatii* from 717094 bp to 891450 bp. These genes encode parallel beta helix repeat containing proteins and display distinctive oligomer patterns. Increasing divergence from average (up to an extreme value at ± 3 standard deviations) is indicated by progressively darker colours. GC content and the three tetranucleotide parameters are plotted on the innermost four rings. Distance (second innermost circle) is the distance between global and local sliding window tetranucleotide patterns, Pattern skew (third innermost circle) is the distance between tetranucleotide rankings on direct and reverse strands, and Oligonucleotide variance (fourth innermost circle) is the numerical variance of oligomers, where a lower value indicates tetramer usage is more highly restricted (for example in repeat regions). Rings 5 (χ^2 threshold 500) and 6 (χ^2 threshold 1200) display the number of bases occupied by overrepresented 8-14mers in a certain region, with overlaps only counted once, as a percentage. The outermost ring shows the difference (in classes) between a tetranucleotide parameter, Oligonucleotide variance, and the 8-14mers in ring 5. Further figures for all Chlorobi are presented in the Online Resource. Figures were created with the program JCircleGraph in the OligoCounter package (Davenport et al. 2009).

Fig. 2 Locally collinear blocks (LCBs) identified among completely sequenced Chlorobi genomes: (a) *C. tepidum*, *Chlorobaculum parvum*, *Chlorobium luteolum* and *Chlorobium chlorochromatii*, and (b) all five *Chlorobium* species and (c) *Chlorobaculum tepidum* and *Chlorobaculum parvum*. Each contiguously colored region is a locally collinear block, a region which does not display rearrangement of homologous backbone sequence. Within each block is a representation of the average level of conservation in that region of sequence. LCBs below the dividing line are in the reverse complement orientation relative to the reference genome. Lines between genomes trace each orthologous LCB through every genome. Each of the blocks has a minimum weight of 4931. The figure was generated by the Mauve rearrangement viewer (Darling et al. 2004).

Fig. 3. Genome comparison BLAST atlas (Hallin et al. 2008) of *C. tepidum* TLS (Eisen et al. 2002).

Each of the concentric circles represent genomic features as described in the figure legend on the right, with the outermost circle corresponding to the top-most feature in the left panel of the legend, and the innermost circle corresponding to the bottom-most feature in their right panel of the legend. Color intensity in the ten outermost circles positively correlates with the degree of sequence homology with the *C. tepidum* ortholog. The following inner eight circles display properties of strain *C. tepidum* TLS. From the 11th to the innermost circle: circles 11, 12: structural features; circles 13, 14: map positions of coding sequences and RNAs on the two strands; circles 15, 16: global repeats; circle 17: GC skew; circle 18: A+T contents.

Fig. 4. Pan-genome gene family clustering, based on 11 *Chlorobi* genome sequences. Numbers represent the bootstrap support for the tree topology with 100 resamplings. The tree was constructed as described in the Methods section.

Fig. 5. Phylogenetic trees of the phylum *Chlorobi*. Both trees were built with MAFFT multiple sequence alignments followed by distance tree construction in JalView with 100 bootstraps. For full details see the Methods section. Top. Phylogeny according to the Fenna-Matthews-Olson (FMO) protein. Bottom. Phylogeny according to the prochlorophyllide reductase iron sulfur ATP-binding protein chlL.

Table 1. Genome characteristics of 11 completely sequenced Chlorobi (green sulfur bacteria) in the NCBI RefSeq collection (July 2009).

Name	RefSeq identifier	Genome size (bp)	GC content (%)	% coding	Number of ORFs	16S rRNA count	Pseudo -genes	CRIS-PRs
<i>Chlorobaculum parvum</i> NCIB 8327	NC_011027	2,289,249	55	86	2141	2	40	1
<i>Chlorobaculum tepidum</i> TLS (= <i>Chlorobium tepidum</i>)	NC_002932	2,154,946	56	87	2337	2	0	2
<i>Chlorobium chlorochromatii</i> CaD3	NC_007514	2,572,079	44	88	2050	1	0	2
<i>Chlorobium limicola</i> DSM245	NC_010803	2,763,181	51	86	2578	2	90	4
<i>Chlorobium luteolum</i> DSM273 (= <i>Pelodictyon luteolum</i>)	NC_007512	2,364,842	57	89	2137	2	0	1
<i>Chlorobium phaeobacteroides</i> BS1	NC_010831	2,736,403	48	82	2619	2	95	4
<i>Chlorobium phaeobacteroides</i> DSM266	NC_008639	3,133,902	48	84	2799	2	93	5
<i>Chlorobium phaeoclathratiforme</i> BU-1 (= <i>Pelodictyon phaeoclathratiforme</i>)	NC_011060	3,018,238	48	85	2979	3	212	3
<i>Chlorobium phaeovibrioides</i> DSM265 (= <i>Prosthecochloris vibrioformis</i>)	NC_009337	1,966,858	52	91	1824	1	20	2
<i>Chloroherpeton thalassium</i> ATCC 35110	NC_011026	3,293,456	45	86	2782	1	22	2
<i>Prosthecochloris aestuarii</i> DSM271	NC_011059	2,512,923	50	87	2388	1	74	5

Table 2. Giant genes (> 15 kb) in *Chlorobi* genomes.

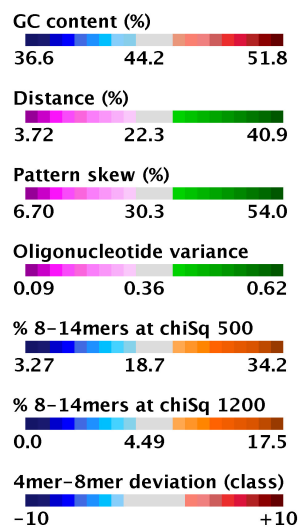
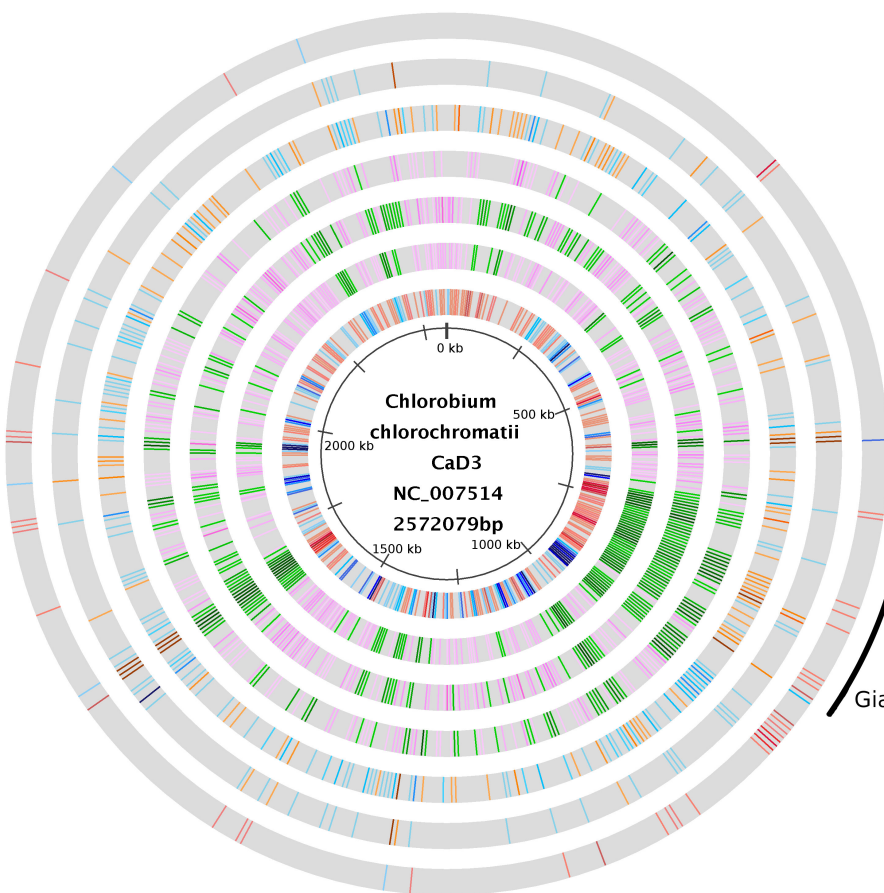
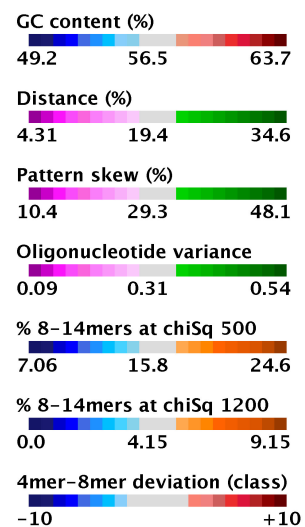
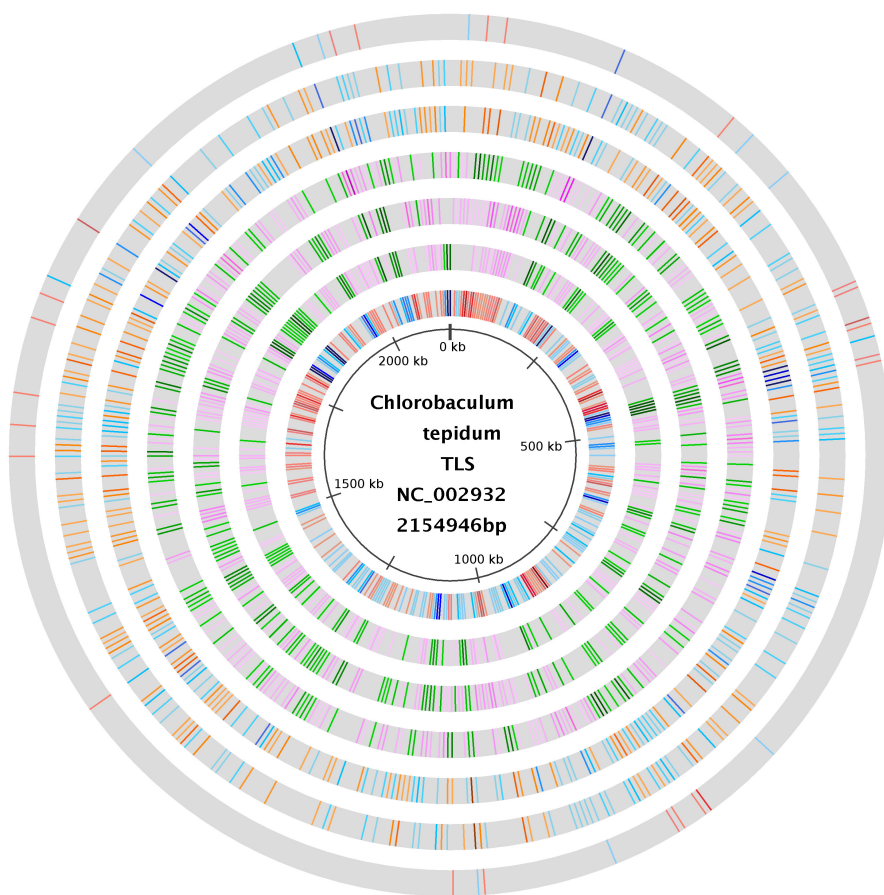
Species	Annotation	Map Start	Position Stop	Size (bp)
<i>Chlorobium chlorochromatii</i> CaD3	Parallel beta-helix repeat	717094	827511	110418
	Parallel beta-helix repeat	829510	891450	61941
	Putative cell adhesion VCBS protein	1656265	1705200	48936
	Putative cell adhesion VCBS protein	1017685	1044300	26616
<i>Chlorobium phaeoclathratiforme</i> BU-1	VCBS putative outer membrane adhesin like protein	610441	632142	21702
<i>Chlorobium phaeovibrioides</i> DSM265	VCBS putative outer membrane adhesin like protein	475816	498102	22287
	VCBS putative outer membrane adhesin like protein	435802	454140	18339
<i>Prosthecochloris aestuarii</i> DSM271	Van Willebrand factor like protein, putative adhesin like protein	828668	843697	15030

Table 3. Anomalous regions in the 11 completely sequenced *Chlorobi*. Regions were selected according to divergence of oligonucleotide parameters from the core genomic average on the SeqWord Genome Browser website (Ganesan et al. 2008). Giant genes are listed in Table 2.

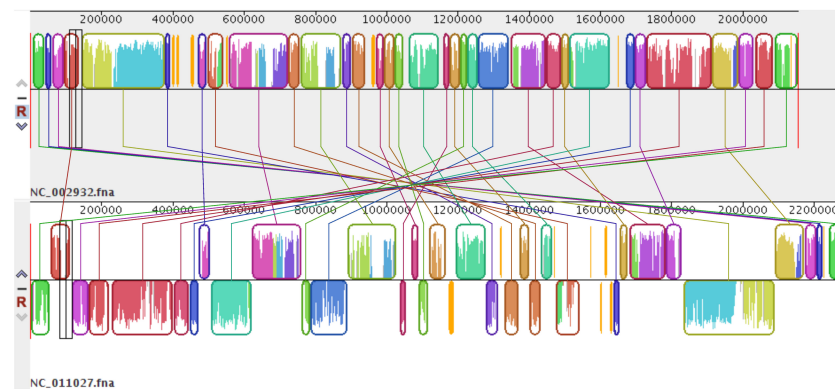
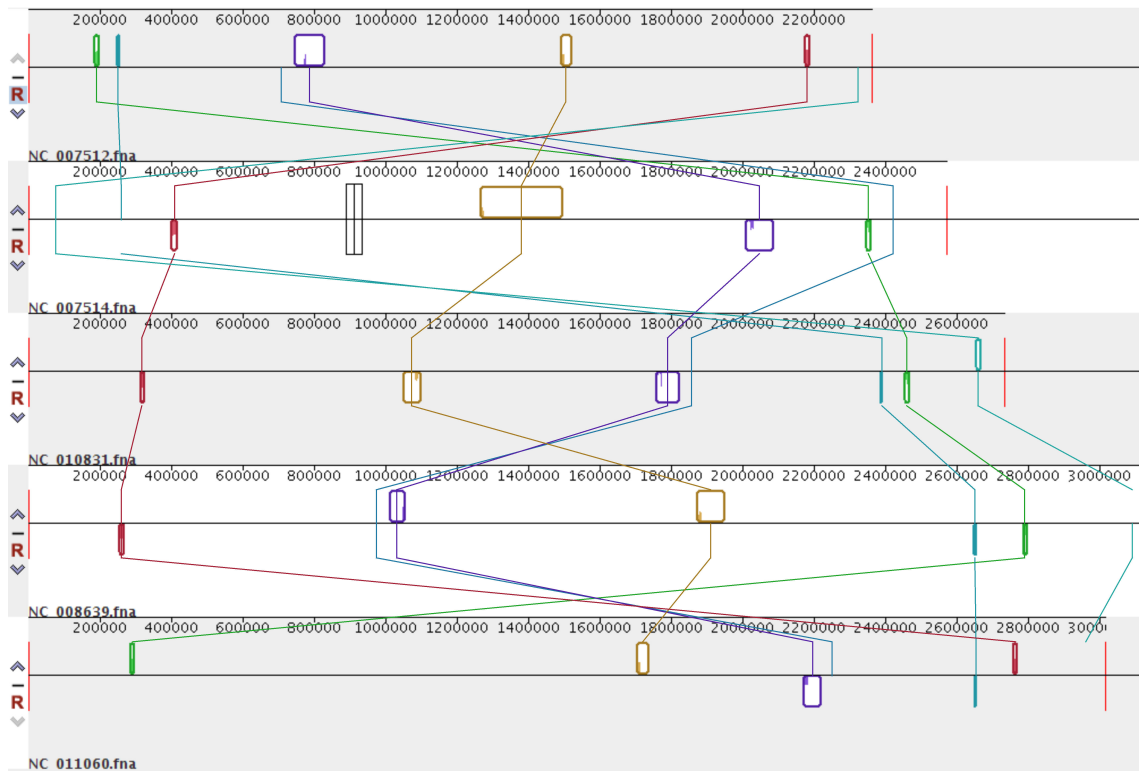
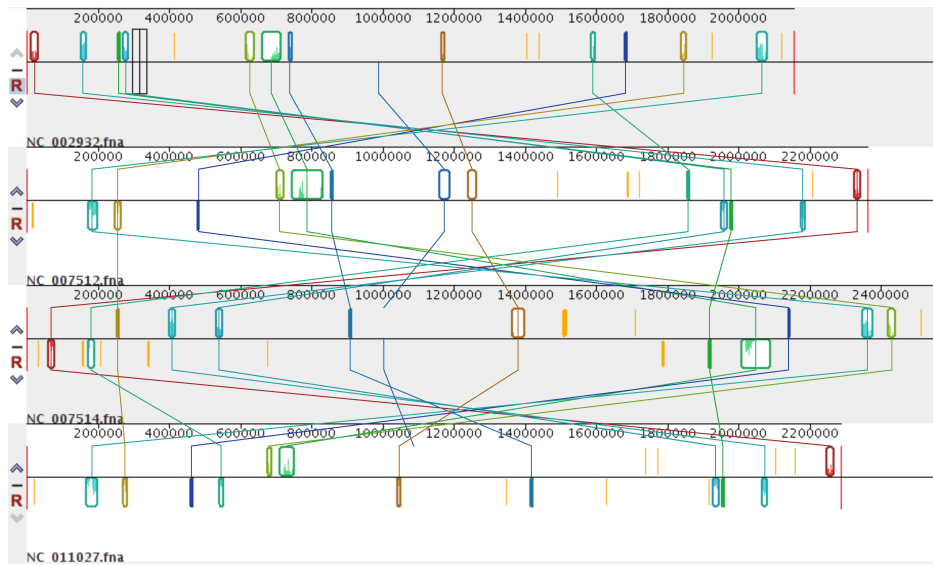
Start coordinate	End coordinate	Annotated genes
		<i>Chlorobaculum parvum</i> NCIB8327
88000	102000	CRISPR proteins, transporters
140000	150000	Synthases, ribonuclease, response regulator
306000	320000	Ribosomal proteins, synthases, ABC transporter, hypotheticals
1030000	1040000	Hydrogenase/sulfur reductase, oxidored, oxidase
1432000	1440000	ABC transporter, sensor kinases
1576000	1596000	Synthases, transferases
1598000	1612000	Type 2/3 secretion system, pili assembly proteins
1982000	2002000	Glycosyl transferases, ABC transporter
		<i>Chlorobaculum tepidum</i> TLS
1856000	1868000	bchE cyclase, enolase, efflux pump, hypotheticals
		<i>Chlorobium chlorochromatii</i> CaD3
338000	346000	Peptidase, hypotheticals
920000	952000	Exopolysaccharide biosynthesis
1050000	1060000	Metalloprotease secretion, type 1 secretion protein TolC
1482000	1490000	SMR drug efflux transporter, put. Regulator, methyltransferase
1848000	1856000	Transcript. Regulator, symporter, endonuclease
1942000	1952000	Glycosyl transferases
		<i>Chlorobium limicola</i> DSM245
250000	262000	IS4 transposase, nucleotidyltransferase
1140000	1148000	Hypotheticals, aldolases
1442000	1454000	Helicases, IS4 transposase, exodeoxyribonuclease
1578000	1590000	Pentapeptide repeat protein, miro domain protein, DNA repair protein, regulators, DNA polymerase beta domain
2082000	2100000	Transposase IS3/IS911, cyclases, aldolases, oxidoreductases, integrase, cytidyltransferase
2480000	2496000	IS4 transposases, ArsR regulator, helicases
		<i>Chlorobium luteolum</i> DSM273
776000	784000	Transferases, transcriptional regulator
800000	808000	gvpA, type I secretion system, metalloprotease secretion protein
884000	892000	Fe-S oxidoreductase-like protein, transferases, hypotheticals
950000	960000	Hypotheticals, DNA-damage-inducible protein J, regulator
1170000	1180000	Type II secretory pathway component, hypotheticals
1464000	1474000	CRISPR-associated proteins, transport system
1554000	1562000	Glycosyl transferases, Fe-S oxidoreductase-like

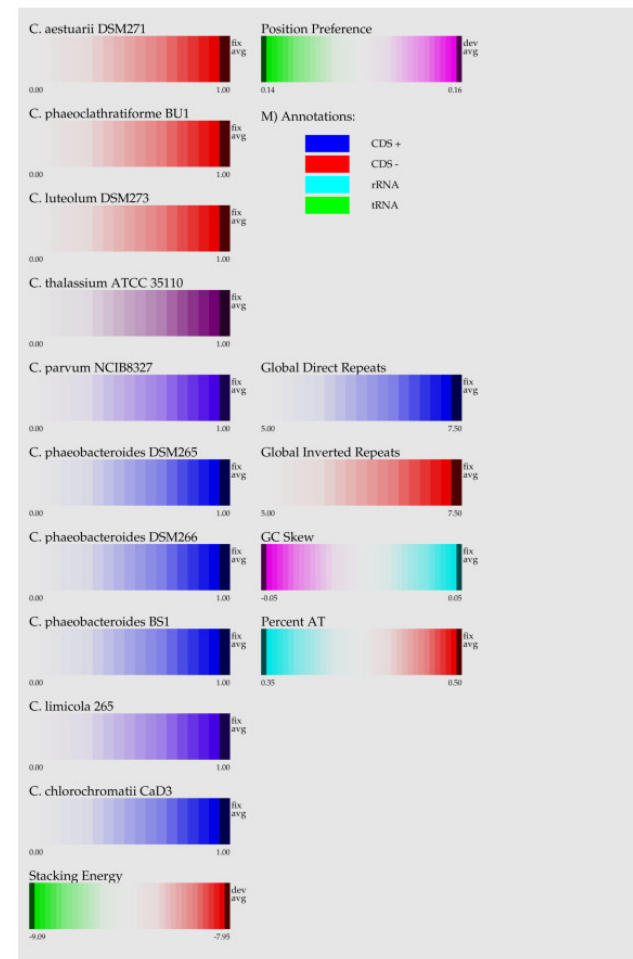
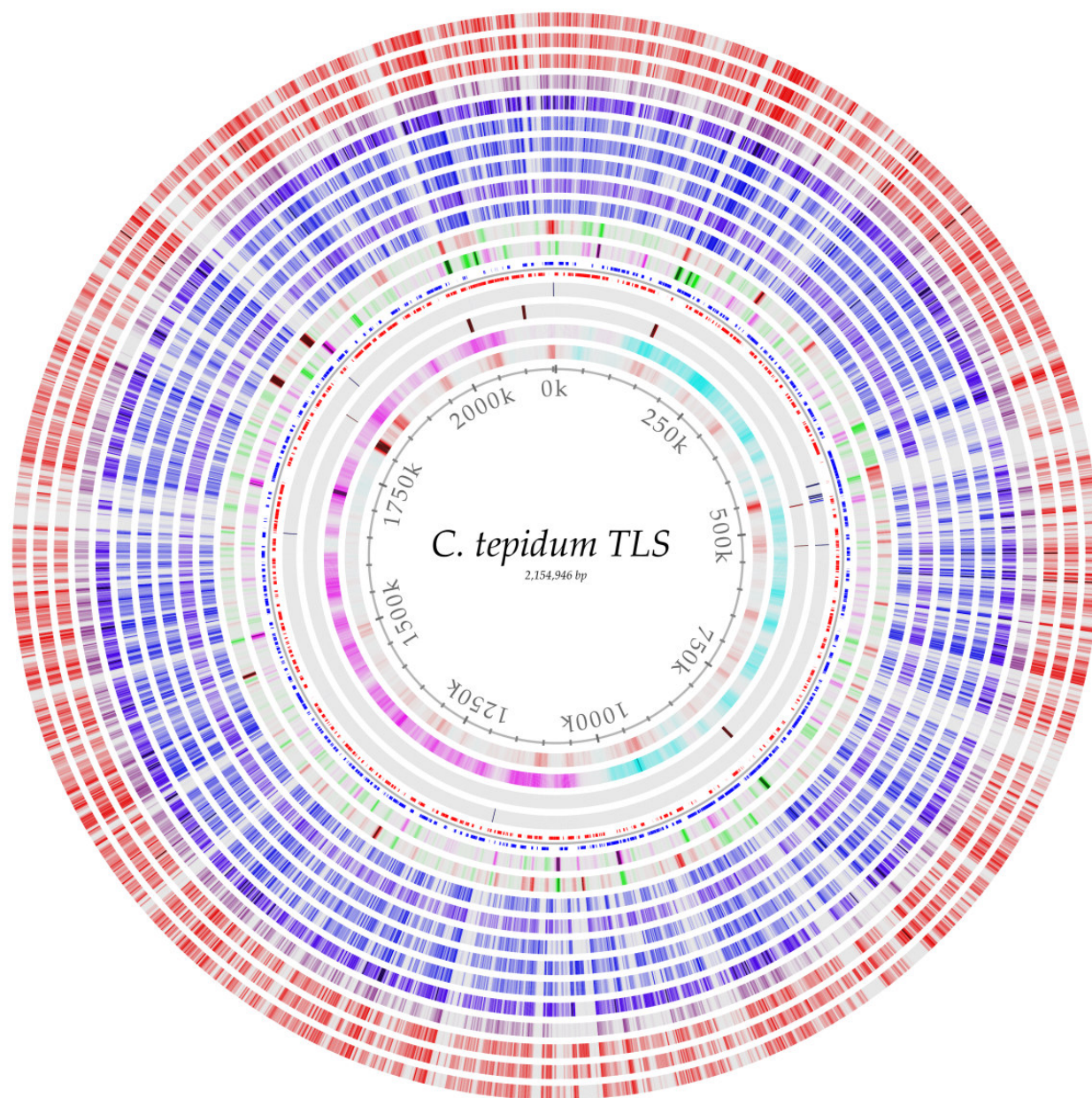
2032000	2042000	RNA-directed DNA polymerase, hypotheticals
2048000	2074000	Polysaccharide biosynthesis protein CapD, regulator, hypotheticals
2078000	2086000	Ligase, endonuclease, O-antigen export
		<i>Chlorobium phaeobacteroides</i> BS1
790000	802000	Two restriction systems, endonuclease, methylase
1234000	1244000	Methylases, type three restriction system, hypotheticals
2086000	2102000	Transposases, biosynthesis, O-antigen polymerase, glycosyl transferase, dehydratases
2474000	2482000	Polysaccharide biosynthesis, phosphotransferases
		<i>Chlorobium phaeobacteroides</i> DSM266
382000	396000	Transposases IS4, ABC transporter, transferases
1472000	1482000	Ankyrin, hypotheticals
1578000	1588000	Metallophosphoesterases, restriction modification system
1832000	1840000	Restriction endonucleases, helicase
2548000	2562000	Transposases, helicases, unknown function
2566000	2588000	Integrase, hypotheticals, ATPase
2708000	2716000	Transposase IS4, transferases, hemolysin-like protein
3000000	3024000	Transposases IS4, polysaccharide biosynthesis, hypotheticals
		<i>Chlorobium phaeoclathratiforme</i> BU-1
182000	190000	PilT domain protein, helicases, Sell repeat protein
496000	504000	Helicases, synthases, cobalamin
520000	532000	LPS biosynthesis, transferases, synthase, oxidoreductases
536000	556000	Polysaccharide biosynthesis protein CapD, transferases, hypotheticals
558000	566000	Transferases, reductases
594000	604000	Transporter, dehydrogenases
606000	614000	Put. adhesin protein, phosphate dikinase
1294000	1304000	Transposase, hypotheticals
1538000	1548000	MM334 transposase, hydrolase
1902000	1920000	Synthases, gas vesicle synthesis, regulators, hypotheticals
1954000	1962000	NADH-quinone oxidoreductase chains
2104000	2122000	Oxidoreductase, ABC transporter, transposases
2530000	2542000	CRISPR associated proteins, reverse transcriptase
2648000	2656000	TonB dependent receptors, magnesium chelatase
		<i>Chloroherpeton thalassium</i> ATCC35110
1200000	1214000	Integrases, transcriptional regulators, PilT, dioxygenases
2278000	2296000	Proton channel, oxidoreductases, symporter, Tetratricopeptide repeat protein
2454000	2464000	Transferases, polysaccharide biosynthesis proteins, aldolases
2538000	2558000	Polysaccharide biosynthesis, metallophosphoesterase, methyltransferases
2564000	2576000	Lyases, transferases, polymyxin resistance protein, reductases

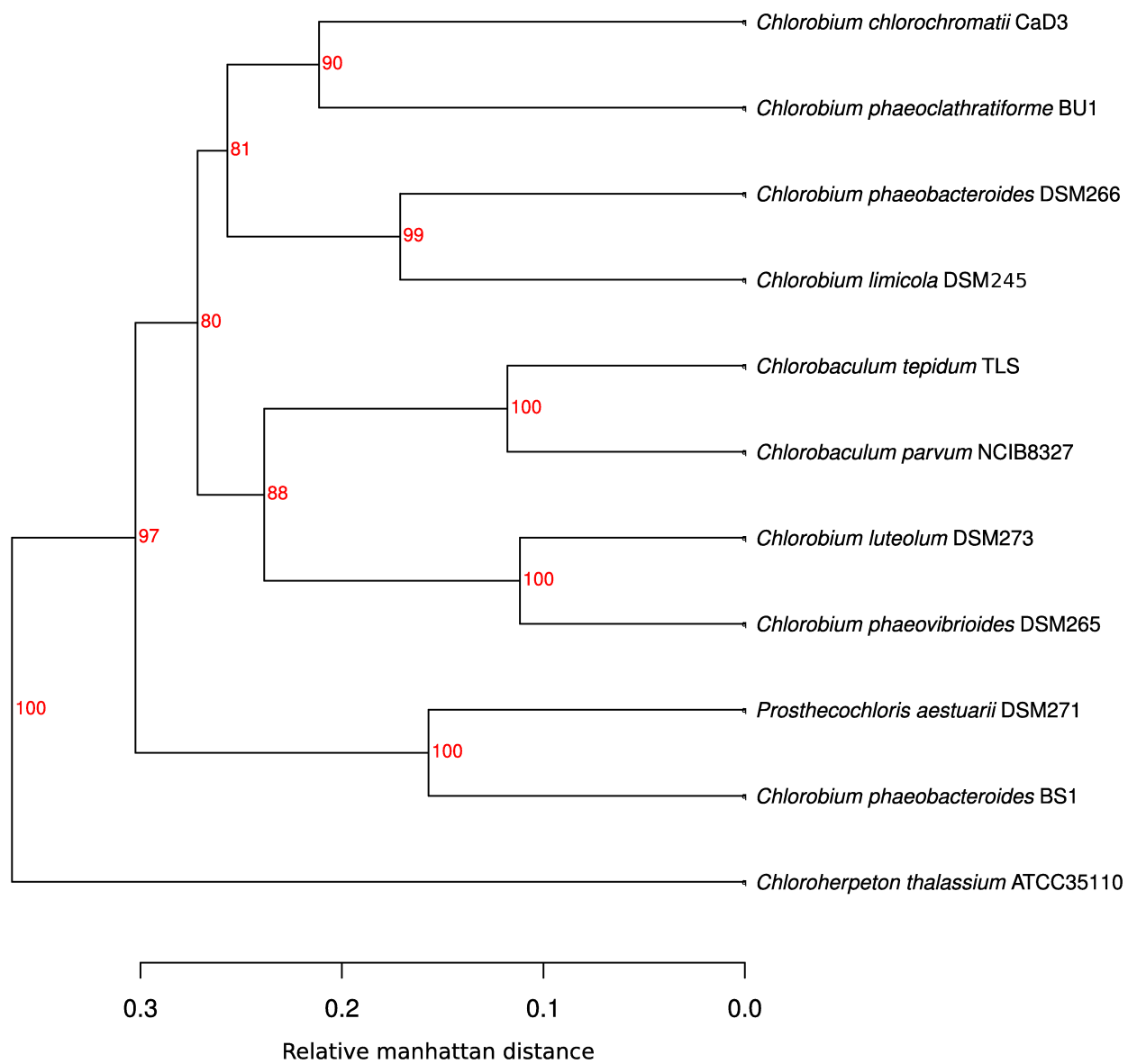
		<i>Prosthecochloris aestuarii</i> DSM271
70000	80000	Transposases, RelA/SpoT, put. restriction endonucleases, DNA polymerase
94000	102000	NADH-Ubiquinone oxidoreductase, antiporter, TonB receptor
674000	686000	Integrase, transposase, hypotheticals
804000	812000	Photosystem P840 reaction center protein PscD, ArsR regulator, oxidoreductases, chaperones
1856000	1880000	Transferases, synthetases, conjugal transfer, type IV transporter
1894000	1916000	pilT, isomerases, kinases, transferases
1918000	1930000	Transferases, polysaccharide biosynthesis proteins
2278000	2290000	Integrase, transposase, hypotheticals
		<i>Prosthecochloris vibrioformis</i> DSM265
101500	106500	Helix-turn-helix domain protein, hypotheticals
137000	142000	Exonuclease, DNA pol. III, regulator, hypotheticals
144500	149500	put. Prophage repressor, DNA polymerase, DNA repair proteins
204000	209000	Synthetases, ribosomal proteins,
520500	525500	Peptidases, exonucleases
707000	712500	Phage integrases, lipopolysaccharide biosynthesis
719000	728000	Transposase ABC transporter, transferases,
732000	741000	Transferases, transporters
742500	757000	Transposases, O-antigen polymerase, transferases, integrase
773500	780500	Transferases, dehalogenases
783000	790000	Transferases, dehydrogenases
1207500	1216500	O-antigen protein, polysaccharide biosynthesis
1357500	1362500	Synthetases, peptidases
1700000	1706500	Surface antigen, porins, synthetases
1717000	1722000	ABC transporter
1848000	1854500	Ferrochelatase, lipoprotein, transporter

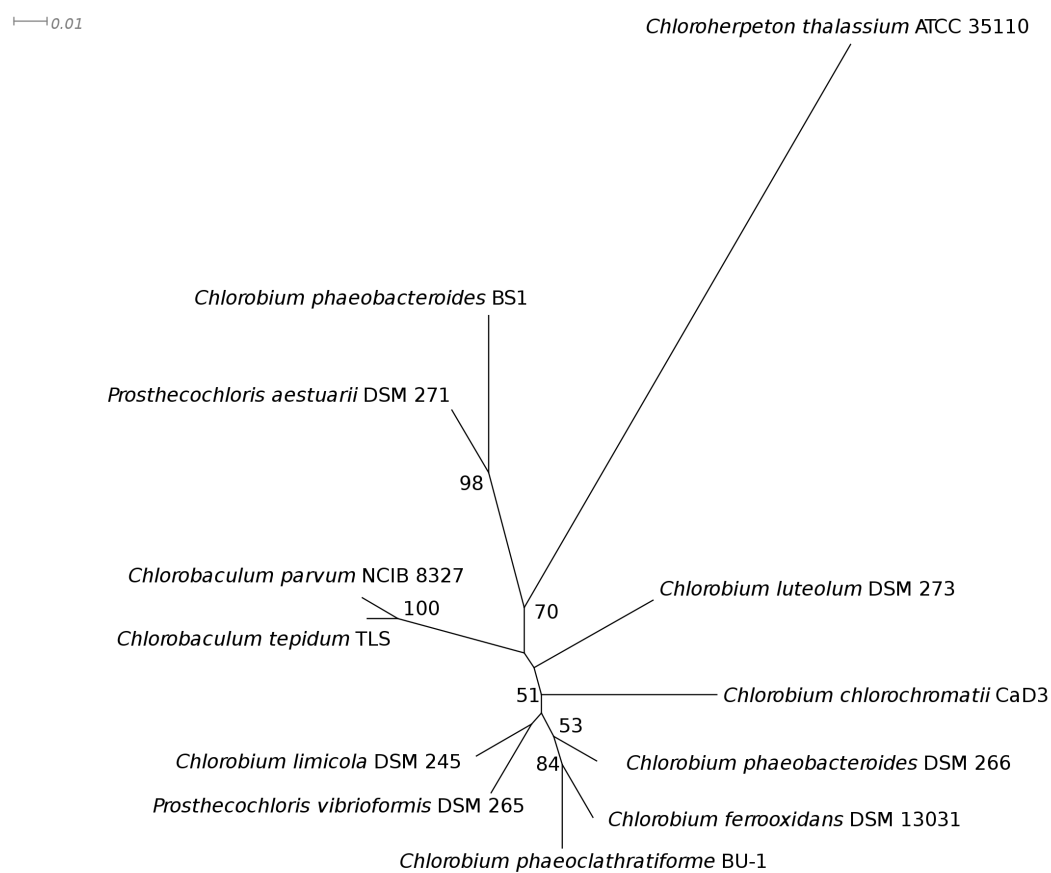
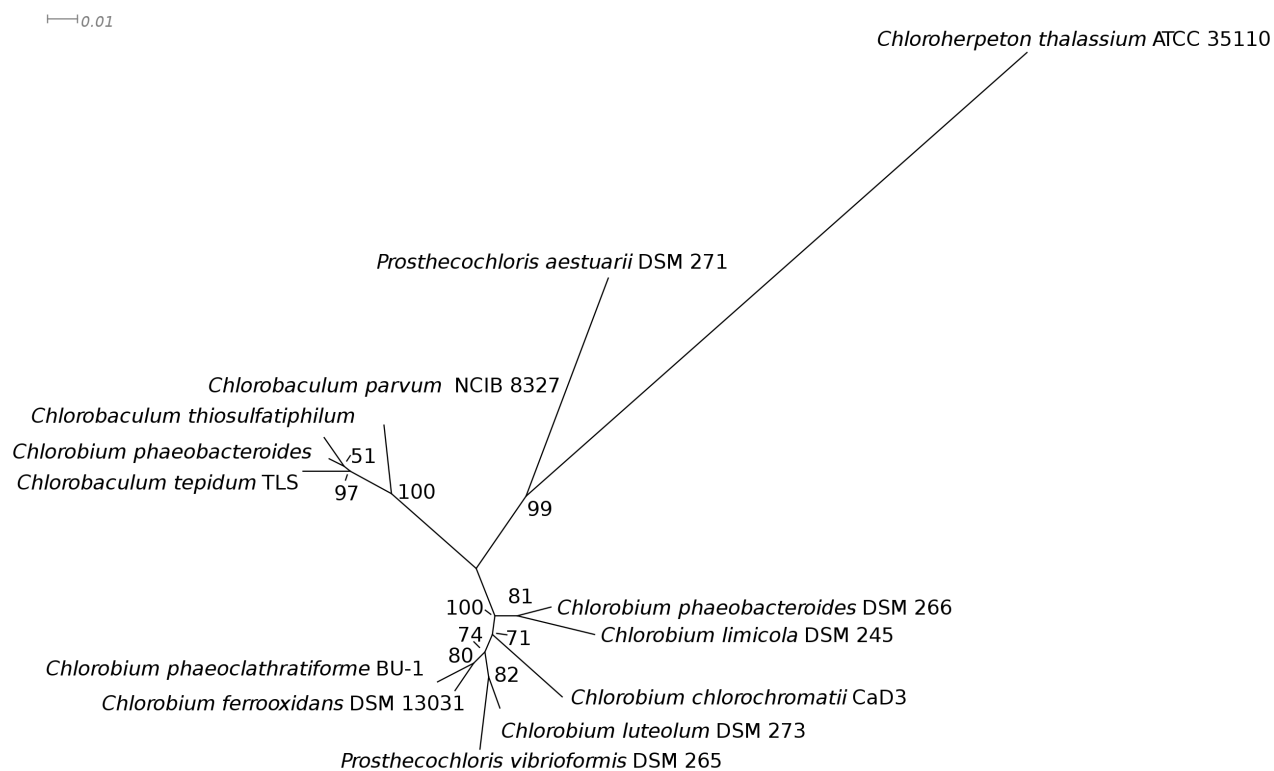


Giant genes









2. The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage

2.1 Background

This part summarises the development and implementation of a web accessible genome browser tool and practical applications of oligonucleotide usage, including examples for multiple genomes. This application was used in the manuscripts listed in section 1.2. Oligonucleotide usage is based on statistical parameters of normalised short oligomers counts in DNA sequences (Karlin et al. 1997). Oligonucleotide usage has been known for some time to be useful in discerning between two genomes which maintain different patterns (Karlin et al. 1997, Abe et al. 2003, Pride et al. 2003, Reva and Tümmler 2004, Teeling et al. 2004). Furthermore, these patterns can also be useful within a genome for discerning between core and accessory genomes (Karlin 2001, Reva and Tümmler 2005, Mrázek et al. 2009). The core genome includes regions where GC content and oligonucleotide patterns are close to the genomic average (Karlin et al. 1997, Reva and Tümmler 2005). In contrast, accessory genomic elements such as integrated plasmids, phages or genomic islands typically divulge different oligomer patterns. Finally, the process of amelioration describes the tendency of monomer or oligomer patterns in blocks of atypical DNA to approach that of the genomic average (Vernikos et al. 2007). This process is difficult to observe or test, but by inference has been linked to replication or repair systems in the genome, and DNA conformational tendencies (Lawrence and Ochman 1997, van Passel et al. 2006). Fascinatingly, plasmids have been demonstrated to maintain a different oligomer signal to the host chromosome (van Passel et al. 2006), perhaps suggesting that other mechanisms are also at work. These might include the frequent uptake of novel DNA by plasmids (van Passel et al. 2006, Cortez et al. 2009), different replication and repair or restriction systems, or the ability of plasmids to be transferred between genomes as part of the "mobilome" (Cortez et al. 2009).

2.2 About the manuscript

The SeqWord Genome Browser was implemented as a Java applet which runs in a web browser and links to a database of precalculated nucleotide patterns. The main achievement of this work was to allow easy access for all researchers to oligonucleotide pattern data for all public completely sequenced prokaryotic genomes. This tool facilitates analysis at multiple genome scales between comparative genomics of complete genomes right down to single genes across several species.

Algorithms had been previously derived and listed in previous work (Reva and Tümmler 2005). The browser applet is flexible, allowing users to browse genomes with annotations in a linear fashion, or alternatively, to view genomic fragments (eg. 10 kbp in size). Fragment view facilitates a novel interpretation of genomes or genes, as the genomic regions may be plotted in three dimensions according to various oligonucleotide parameters. As oligonucleotide parameters contain structural and/or phylogenetic signal, conserved and accessory genomic regions can be differentiated. For example, the small chromosome of *Vibrio cholerae* El Tor N16961 has large regions with highly variable genomic content. This fascinating replicon harbours an integron island termed a "gene capture system" and contains multiple regions of putatively horizontally acquired DNA, even from outside the gammaproteobacteria (Heidelberg et al. 2000). It seems the large scale integration of novel DNA into this chromosome is sufficient to disrupt its global oligonucleotide usage. In other words, according to oligonucleotide usage this chromosome does not consist of well defined core and accessory regions, but is a true mosaic of genetic elements from diverse origins (Heidelberg et al. 2000). Interestingly, the larger chromosome of *V. cholerae* displays considerably different oligonucleotide usage patterns, and the small chromosome has even suggested to be a captured megaplasmid (Heidelberg et al. 2000). Another use case on single genes is the absence of any hallmarks of divergent oligonucleotide usage in the *narG* gene (Palmer et al. 2009). Furthermore, classes of genes such as long modular genes, ribosomal RNA or ribosomal proteins with distinct oligonucleotide properties could be located. This web service thus allows tentative assignment of outlier novel fragments to a few classes of genes. SeqWord is available online from sites at Pretoria, South Africa and Hanover, Germany.

The SeqWord Genome Browser was published in BMC Bioinformatics in 2008. I was mainly involved with writing the manuscript, assessing the scientific content of the system and maintaining the mirror site server in Hanover. Oleg Reva was the architect of the whole project, while Anna Rakitianskaia and Hamilton Ganesan did the programming and technical aspects. Burkhard Tümmler had previously developed the statistical parameters used in this work.

2.3 References

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. & Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res* **13**(4) 693-702.

- Cortez, D., Forterre, P. & Gribaldo, S. (2009) A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol* **10**(6) R65.
- Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., et al. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**(6795) 477-483.
- Karlin, S., Mrázek, J. & Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12) 3899-3913.
- Karlin, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9**(7) 335-343.
- Lawrence, J. G. & Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**(4) 383-397.
- Mrázek, J. (2009) Phylogenetic signals in DNA composition: limitations and prospects. *Mol Biol Evol* **26**(5) 1163-1169.
- Palmer, K., Drake, H. L. & Horn, M. A. (2009) Genome-derived criteria for assigning environmental narG and nosZ sequences to operational taxonomic units of nitrate reducers. *Appl Environ Microbiol* **75**(15) 5170-5174.
- van Passel, M., Bart, A., Luyf, A., van Kampen, A. & van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7**(1) 26.
- Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2) 145-158.
- Reva, O. N. & Tümmler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**, 90.
- Reva, O. N. & Tümmler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163.
- Vernikos, G. S., Thomson, N. R. & Parkhill, J. (2007) Genetic flux over time in the *Salmonella* lineage. *Genome Biol* **8**(6) R100.

Software

Open Access

The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage

Hamilton Ganesan¹, Anna S Rakitianskaia¹, Colin F Davenport², Burkhard Tümmler² and Oleg N Reva^{*1}

Address: ¹Dep. of Biochemistry, Bioinformatics and Computational Biology Unit, University of Pretoria, Lynnwood road, Hillcrest, Pretoria, 0002, South Africa and ²Klinische Forschergruppe, OE 6710, Medizinische Hochschule Hannover, D-30625 Hanover, Germany

Email: Hamilton Ganesan - ham.kzn@gmail.com; Anna S Rakitianskaia - annar@cs.up.ac.za; Colin F Davenport - davenport.colin@mh-hannover.de; Burkhard Tümmler - tuemmler.burkhard@mh-hannover.de; Oleg N Reva^{*} - oleg.reva@up.ac.za

^{*} Corresponding author

Published: 7 August 2008

Received: 29 March 2008

BMC Bioinformatics 2008, 9:333 doi:10.1186/1471-2105-9-333

Accepted: 7 August 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/333>

© 2008 Ganesan et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Data mining in large DNA sequences is a major challenge in microbial genomics and bioinformatics. Oligonucleotide usage (OU) patterns provide a wealth of information for large scale sequence analysis and visualization. The purpose of this research was to make OU statistical analysis available as a novel web-based tool for functional genomics and annotation. The tool is also available as a downloadable package.

Results: The SeqWord Genome Browser (SWGB) was developed to visualize the natural compositional variation of DNA sequences. The applet is also used for identification of divergent genomic regions both in annotated sequences of bacterial chromosomes, plasmids, phages and viruses, and in raw DNA sequences prior to annotation by comparing local and global OU patterns. The applet allows fast and reliable identification of clusters of horizontally transferred genomic islands, large multi-domain genes and genes for ribosomal RNA. Within the majority of genomic fragments (also termed genomic core sequence), regions enriched with housekeeping genes, ribosomal proteins and the regions rich in pseudogenes or genetic vestiges may be contrasted.

Conclusion: The SWGB applet presents a range of comprehensive OU statistical parameters calculated for a range of bacterial species, plasmids and phages. It is available on the Internet at <http://www.bi.up.ac.za/SeqWord/mhhapplet.php>.

Background

The study of genome OU signatures has a long history dating back to early publications by Karlin *et al.* who focused mainly on dinucleotide compositional biases and their evolutionary implications [1-3]. Statistical approaches of OU comparison were further advanced by Deschavanne *et al.*, who applied chaos game algorithms [4]; and by Pride

et al., who extended the analysis to tetranucleotides using Markov Chain Model simulations [5]. Later, a number of practical tools for phylogenetic comparison of bacterial genomes [4,6,7], identification of horizontally transferred genomic islands [8-13] and assignment of unknown genomic sequences [14,15] based on OU statistics became publicly available. These approaches exploited the notion

that genomic OU composition was less variable within genomes rather than between them, regardless of which genomic regions had been taken into consideration [16]. A general belief was that if a significant compositional difference was discovered in genomic fragments relative to the core genome, these loci most likely can be assigned to horizontally transferred genetic elements (transposons, prophages or integrated plasmids). This approach was criticized by several researchers [17,18] who pointed out that codon bias and base composition are poor indicators of horizontal gene transfer. Therefore, there is a need for more informative parameters which also take into account higher order DNA variation. An overview of the current OU statistical methods based on di-, tetra- and hexanucleotides has been published recently. The conclusion of the review was that all methods were context dependent and, though being efficient and powerful, none of them were superior in all applications [19]. Thus, the major motivation of our work was to develop more flexible and informative algorithms seamlessly integrating di- to heptanucleotides OU analysis for reliable identification of divergent genomic regions.

Recently we have introduced the concept of OU patterns into the literature [20]. Each OU pattern is characterized by a number of OU statistical parameters namely, local pattern deviation (D), pattern skew (PS), relative variance (RV) and others (see Methods section). Novelty of the developed algorithms relative to other existing methods include the following: i) distances between patterns of different word length (from di- through to heptanucleotides) calculated for the same sequences are comparable; i.e. one may use longer word patterns to perform a large scale analysis and then switch to shorter word patterns for a more detailed view; ii) OU patterns calculated for sequences of different lengths are comparable provided that the length of the sequence is longer than the corresponding thresholds (specified in the Methods section); iii) alterations of OU patterns may be analyzed by different non-redundant parameters (D, PS and RV with different schemes of normalization by frequencies of shorter constituent words). Superimposition of these OU characteristics allows better discrimination of divergent genomic regions relative to other contemporary approaches [21].

Implementation

Calculation of OU statistical parameters has been described previously [20,21]. OU pattern was denoted as a matrix of deviations $\Delta_{[\xi_1 \dots \xi_N]}$ of observed from expected counts for all possible words of length N :

$$\Delta_{[\xi_1 \dots \xi_N]} = (C_{[\xi_1 \dots \xi_N]|obs} - C_{[\xi_1 \dots \xi_N]|e}) / C_{[\xi_1 \dots \xi_N]|0} \quad (1)$$

where ξ_n is any nucleotide A, T, G or C in the N -long word; $C_{[\xi_1 \dots \xi_N]|obs}$ is the observed count of the word $[\xi_1 \dots \xi_N]$; $C_{[\xi_1 \dots \xi_N]|e}$

is the expected count and $C_{[\xi_1 \dots \xi_N]|0}$ is a standard count estimated from the assumption of an equal distribution of words in the sequence: $(C_{[\xi_1 \dots \xi_N]|0} = L_{seq} \times 4^{-N})$.

Expected counts of words $C_{[\xi_1 \dots \xi_N]|e}$ were calculated in accordance with the applied normalization scheme. Thus, $C_{[\xi_1 \dots \xi_N]|e} = C_{[\xi_1 \dots \xi_N]|0}$ if OU is not normalized, or $C_{[\xi_1 \dots \xi_N]|e} = C_{[\xi_1 \dots \xi_N]|n}$ if OU is normalized by empirical frequencies of all shorter words of the length n . The expected count of a word $C_{[\xi_1 \dots \xi_N]|e}$ of length N in a L_{seq} long sequence normalized by frequencies of n -mers ($n < N$) was calculated as follows:

$$C_{[\xi_1 \dots \xi_N]|n} = L_{seq} \times F_{[\xi_1 \dots \xi_n]} \times \prod_{i=2}^{N-n+1} \left(\frac{F_{[\xi_i \dots \xi_{i+n-1}] \xi_{i+n}}}{\sum_{\xi} \frac{F_{[\xi_i \dots \xi_{i+n-1}] \xi}}{A, T, G, C}}} \right) \quad (2)$$

where the $F_{[\xi_1 \dots \xi_n]}$ values are the observed frequencies of the particular word of length n in the sequence and ξ is any nucleotide A, T, G or C. For example, expected count of a word ATGC in a sequence of L_{seq} nucleotides normalized by frequencies of trinucleotides is:

$$C_{ATGC} = L_{seq} \times F_{ATG} \times \frac{F_{TGC}}{F_{TGA} + F_{TGT} + F_{TGG} + F_{TGC}} \quad (3)$$

Two approaches of normalization have been exploited where the F values were calculated for the complete sequence of a chromosome, plasmid, etc (generalized normalization) or for a given sliding window (local normalization). The normalization by equation 2 allows identification of words, frequencies of which cannot be predicted exactly by frequencies of shorter constituent words.

The distance D between two patterns was calculated as the sum of absolute distances between ranks of identical words (w , in a total 4^N different words) after ordering of words by $\Delta_{[\xi_1 \dots \xi_N]}$ values (see equation 1) in patterns i and j as follows:

$$D(\%) = 100 \times \frac{\sum_{w=1}^{4^N} |rank_{w,i} - rank_{w,j}| - D_{min}}{D_{max} - D_{min}} \quad (4)$$

Application of ranks instead of relative oligonucleotide frequency statistics made the comparison of OU patterns less biased to the sequence length provided that the sequences are longer than the limits of 0.3, 1.2, 5, 18.5, 74

and 295 kbp for di-, tri-, tetra-, penta-, hexa- and heptanucleotides, respectively [20].

PS is a particular case of D where patterns i and j were calculated for the same DNA but for direct and reversed strands, respectively. $D_{\max} = 4^N \times (4^N - 1)/2$ and $D_{\min} = 0$ when calculating a D or, in a case of PS calculation, $D_{\min} = 4^N$ if N is an odd number or $D_{\min} = 4^N - 2^N$ if N is an even number due to presence of palindromic words [20]. Normalization of D-values by $D_{\max}$ ensures that the distances between two sequences are comparable regardless of the word length of OU patterns.

Relative variance of an OU pattern was calculated by the following equation:

$$RV = \frac{\sum_w \Delta_w^2}{\left(4^N - 1\right) \sigma_0^2} \quad (5)$$

where N is word length; Δ_w^2 is the square of a word w count deviation (see equation 1); and σ_0^2 is the expected variance of the word distribution in a randomly generated sequence that depends on the sequence length and the word length:

$$\sigma_0^2 = 0.14 + \frac{4^N}{L_{seq}} \quad (6)$$

where L_{seq} is sequence length, and N is word length. Normalization of OU pattern variance by σ_0 makes the variances comparable regardless of the word length of OU patterns and the sequence length. The regression equation was tested on 300 randomly generated sequences with an equiprobable occurrence of all 4 nucleotides by the DataFit 7.1.44 software.

The SWGB is coded in Java to be used as an applet in a Web-browser either on the Internet or locally (the programs OligoWords in Python and SeqWord_Viewer, which respectively calculate and visualize the OU patterns for DNA sequences, are available for download from the

SWGB website). SWGB should run on any platform with a Java 1.5.x runtime environment or newer.

The pre-calculated data-sets are saved in a MySQL Server 5.0 database. The size of the sliding window and the OU pattern type were applied according to the sequence length (Table 1). At the time of writing, the SeqWord database contained OU patterns pre-calculated for the sequences of 682 bacterial chromosomes belonging to 637 different organisms (strains and species), 412 plasmids, 100 bacteriophages and 39 other viruses, which were downloaded from the NCBI [44].

Results

User familiarity with the abbreviations of the various OU statistical parameters is important. Different types of OU patterns were abbreviated as type_Nmer. Types might be "n0" for non-normalized, or "n1" for normalized by mononucleotide frequencies. For example, the non-normalized tetranucleotide usage pattern is denoted as n0_4mer; tetranucleotide usage pattern normalized by mononucleotide content is n1_4mer etc. The genomes in the SWGB database were analyzed by the following statistical parameters: D – distance between two patterns of the same type (in this work we used distances D between local patterns calculated for overlapping genome fragments and the global genome patterns calculated for the complete sequence – the local pattern deviation); PS – pattern skew, distance between the two patterns of the direct and reverse strands of the same DNA sequence; RV and GRV – oligonucleotide usage variances normalized locally and globally, respectively, and reduced to the OU variance expected for a randomly generated sequence (see Materials section); GC-content (GC) and GC-skew (GCS) in DNA fragments.

The applet GUI and database of pre-calculated OU patterns

The SeqWord Genome Browser (SWGB) applet is available via the Internet [22-24] and is mouse and menu driven. The Web-based applet is used to visualize DNA compositional variations in bacterial and viral genomes stored in the SeqWord database. Every genome in the database is represented by a set of statistical OU parameters (D, PS, GV, GRV, GC and GCS) calculated for genomic fragments, which were selected by a sliding win-

Table 1: Sliding window size and OU pattern types (oligomer lengths) selected for sequences of different length present in the SeqWord database.

Sequence length	Sliding window	Step	OU pattern type
> 2 Mbp	8 kbp	2 kbp	4 mer
from 1 mbp to 2 Mbp	5 kbp	0.5 kbp	4 mer
from 0.5 mbp to 1 Mbp	3 kbp	0.3 kbp	3 mer
< 0.5 Mbp	1.5 kbp	0.15 kbp	3 mer

dow (sliding window length and step were set according to the total length of the sequence as demonstrated in Table 1). While in 70 to 99% of genomic fragments the OU compositional bias is similar to the complete genome OU pattern, some regions with atypical OU composition, however, are always present. Superimposition of different OU parameters allows discrimination of divergent genomic regions, as was published previously [21]. Briefly: rRNA operons are characterized by extremely high PS and low RV; giant genes with multiple repeated elements have high or moderate PS and high RV; horizontally transferred genetic elements are characterized by increased divergence between RV and GRV accompanied by high D; and genes for ribosomal proteins show a moderate increase of D, PS and RV above genomic averages. Having analyzed 1243 sequences of different microorganisms including viruses and plasmids in the SeqWord database, we confirmed that the approaches we have developed and tested previously [25] (mainly on *Pseudomonas putida* KT2440 chromosomal DNA) are appropriate and useful for analysis of genomic sequences of other microorganisms and viruses.

In an open applet window, the user has the ability to choose from an ever growing list of available sequences (Fig. 1). The user also has the option of restricting the list to display only bacterial chromosomes, plasmids, phages, viruses or all sequences by selecting the corresponding filter button. Users have to select a genome in the list and click the 'Display in the Applet' button to retrieve the pre-calculated data. All OU parameters calculated for a given genome may be exported to a local text file by using the 'Export' function from the applet's 'File' menu. Later, instead of again having to connect to the database, users may open and view their local files (previously exported from the applet or calculated by the OligoWords program, see below) via the 'Open' function in the 'File' menu.

The SWGB is basically comprised of two views, denoted by the 'Gene Map' and 'Diagram' tabs. The applet is instrumental for visualization of natural variation in DNA sequences by the interactive diagrams on the 'Gene Map' and 'Diagram' tabs. Users may save the current diagram in JPG format by using the 'Save picture' function in the 'File' menu.

The 'Gene Map' tab offers a simple view of an entire genome at a glance and gives users access to a number of important pre-calculated OU statistics superimposed on the gene map (Fig. 2). Displays for each of the statistical parameters can be toggled on/off by checking items in the 'Hide Elements' menu. By merely mousing over any region on the plot, a message displaying detailed information for the pointed curve will be shown in the 'Message'

bar. Clicking a gene on the map displays a dialog with the annotation details (Fig. 2).

The 'Zoom' function is straight-forward and allows users to control the amount of data viewed in the plot area. Clicking the 'Enter' button after setting the desired zoom value will then redraw the map. A 'Zoom into region' function under the 'Tools' drop-down menu allows users to zoom into exact genomic regions by merely entering their desired co-ordinates into the pop-up dialog box. The 'Tools' → 'Select region' menu item allows highlighting of selected regions without zooming. Use the option 'Clear ...' in the 'Tools' menu to undo zooming or highlighting. To locate a genomic region by gene, click the button 'Select Gene'. In the pop-up dialog box one may order the gene list by gene names, functionality or coordinates, then select a gene in the list and click 'OK'. When a gene annotation is not available, the values of the locus coordinates are used as a gene name. The applet window will be scrolled to the selected gene highlighted on the map (see Fig. 2).

The 'Diagram' tab allows flexible filtering of the underlying data based on the criteria chosen by users. Although the underlying data is pre-calculated, the user may, by simply changing selected parameters, generate very different images which give different insights into the natural genomic variation. To start with, the 'Diagram' view offers a bar chart or a dot-plot presentation of the pre-calculated data. To view a bar chart of the distribution statistics for a given OU parameter, select the desired parameters from the X or Y-axis drop-downs and click 'Enter'. The number of bars displayed can be adjusted using the '# Bars' selector.

On the dot-plot diagram, each genomic fragment (selected by the sliding window) is represented by a dot with X and Y coordinates that correspond to values of OU parameters chosen from X and Y drop-down lists, respectively. The Z axis parameter may be set as well. In this case, the dots are coloured by values of OU parameters selected for the Z axis, and the colour range is displayed on the vertical colour bar on the left of the plot area (Fig. 3).

Having set up the dot-plot, users will be able to identify divergent genomic regions (see next section). To retrieve annotations of genomic fragments corresponding to a group of dots, frame the dots of interest by clicking and dragging over the desired area. A selector frame then appears around the dots (Fig. 3). Clicking the 'Get' button displays the selected genomic fragments with their coordinates and gene annotations. Furthermore, identification and isolation of specific genomic regions may be improved significantly by filtering dots by OU parameters. The simplest way of filtering is by the third (Z axis)

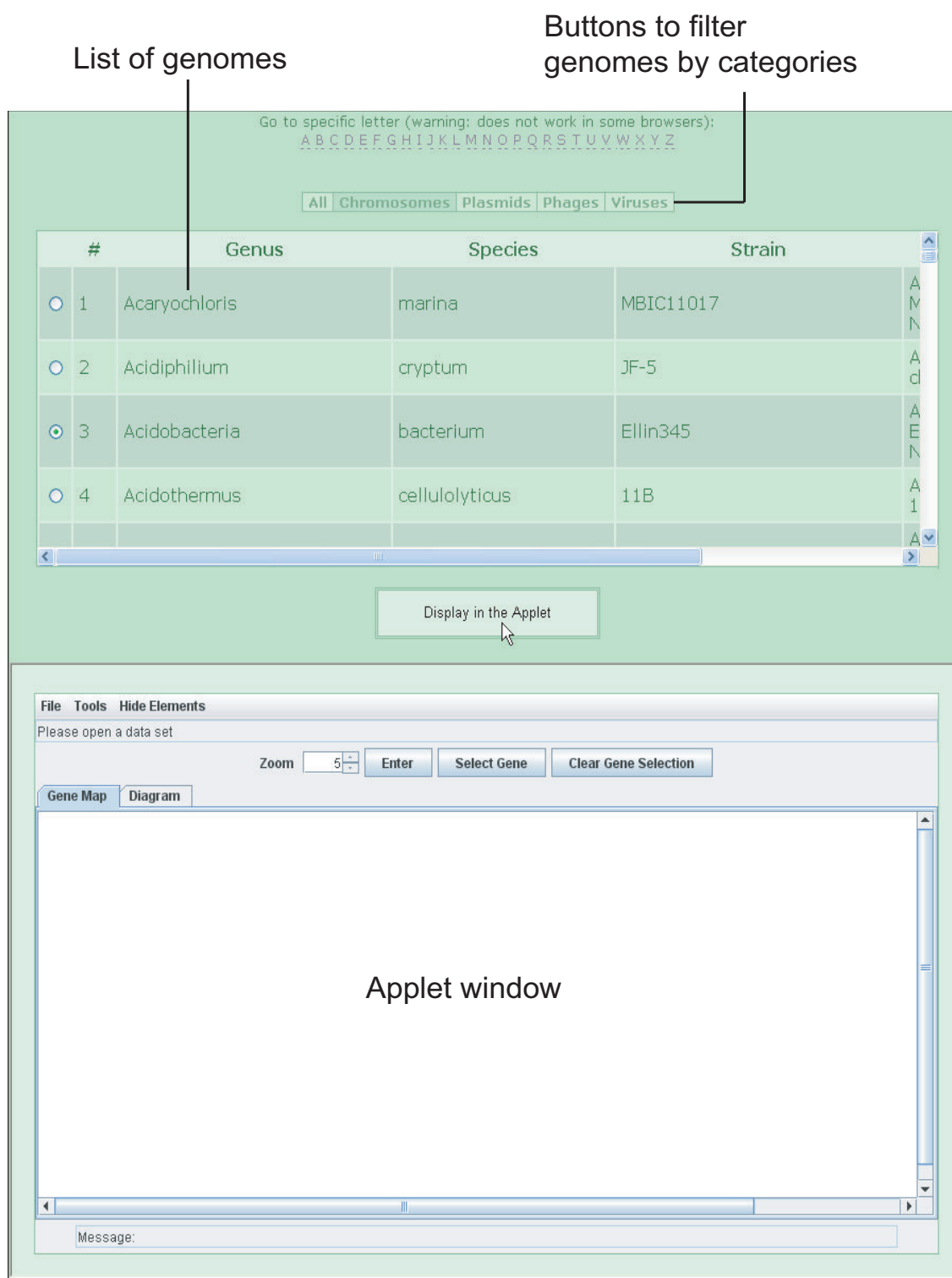


Figure 1
General view of the web-based SWGB with a list of genomes present in the database and an enclosed Java applet for data visualization. To show OU statistical parameters for a selected genome, click the 'Display in the Applet' button. Click a filter button to order genomes by the corresponding category and use the interactive letters at the top to scroll the list to a sequence of interest.

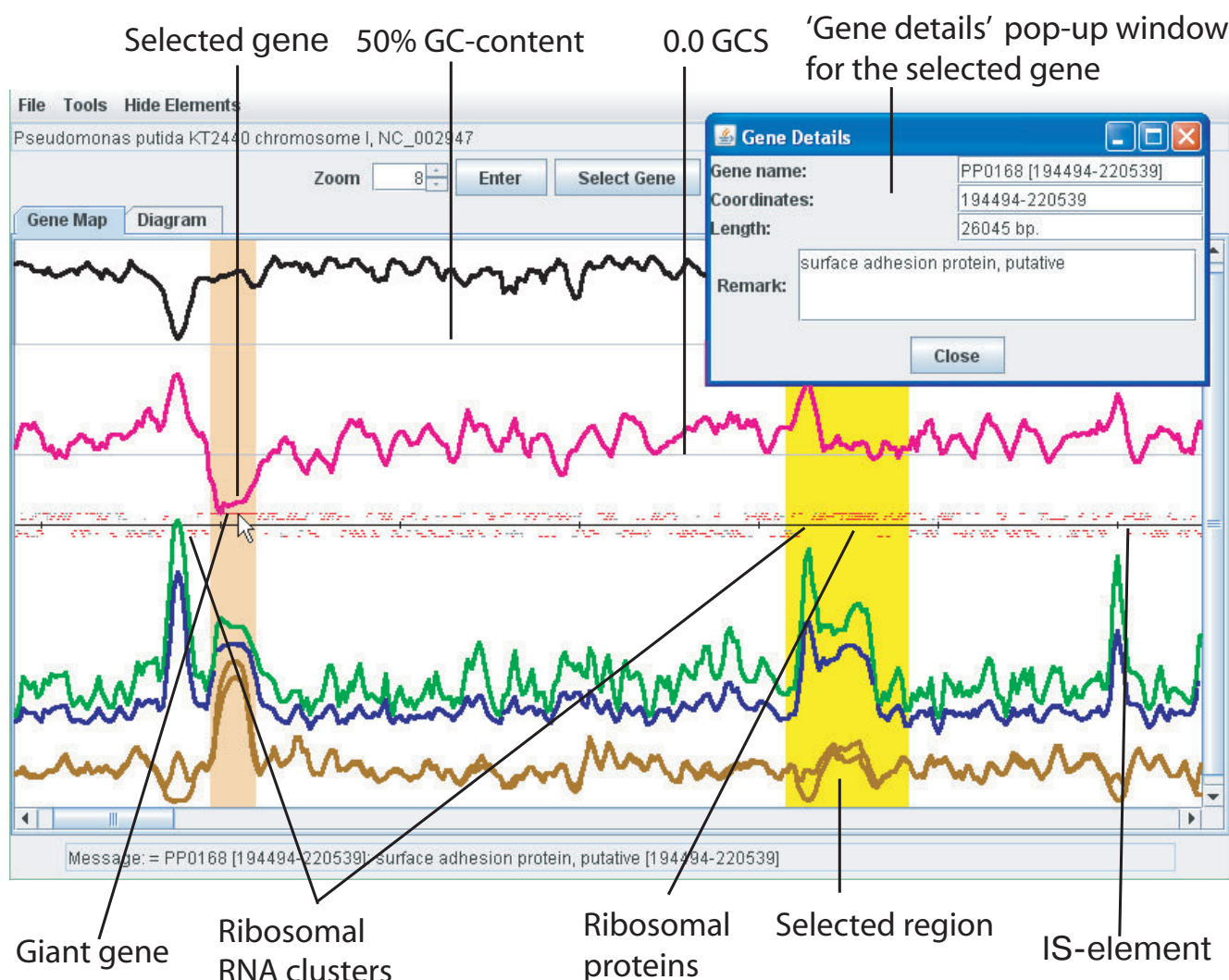


Figure 2

Identification of divergent genomic regions on the 'Gene Map' view. Superimposition of different OU parameters such as GC (black line), GCS (pink), PS (green), D (blue), GRV (upper brown line) and RV (lower brown line) allows discrimination of divergent genomic regions. In this example a part of the chromosome of *Pseudomonas putida* KT2440 (127–774 kbp) is displayed in the applet window. A genomic fragment was highlighted using the function 'Select region' and a giant gene, PP0168, was selected by 'Select gene'. A pop-up window 'Gene Details' was opened by double-clicking the gene on the map. Genes are indicated by red and grey (for hypotheticals) bars. The black horizontal line separates genes by their direction of translation.

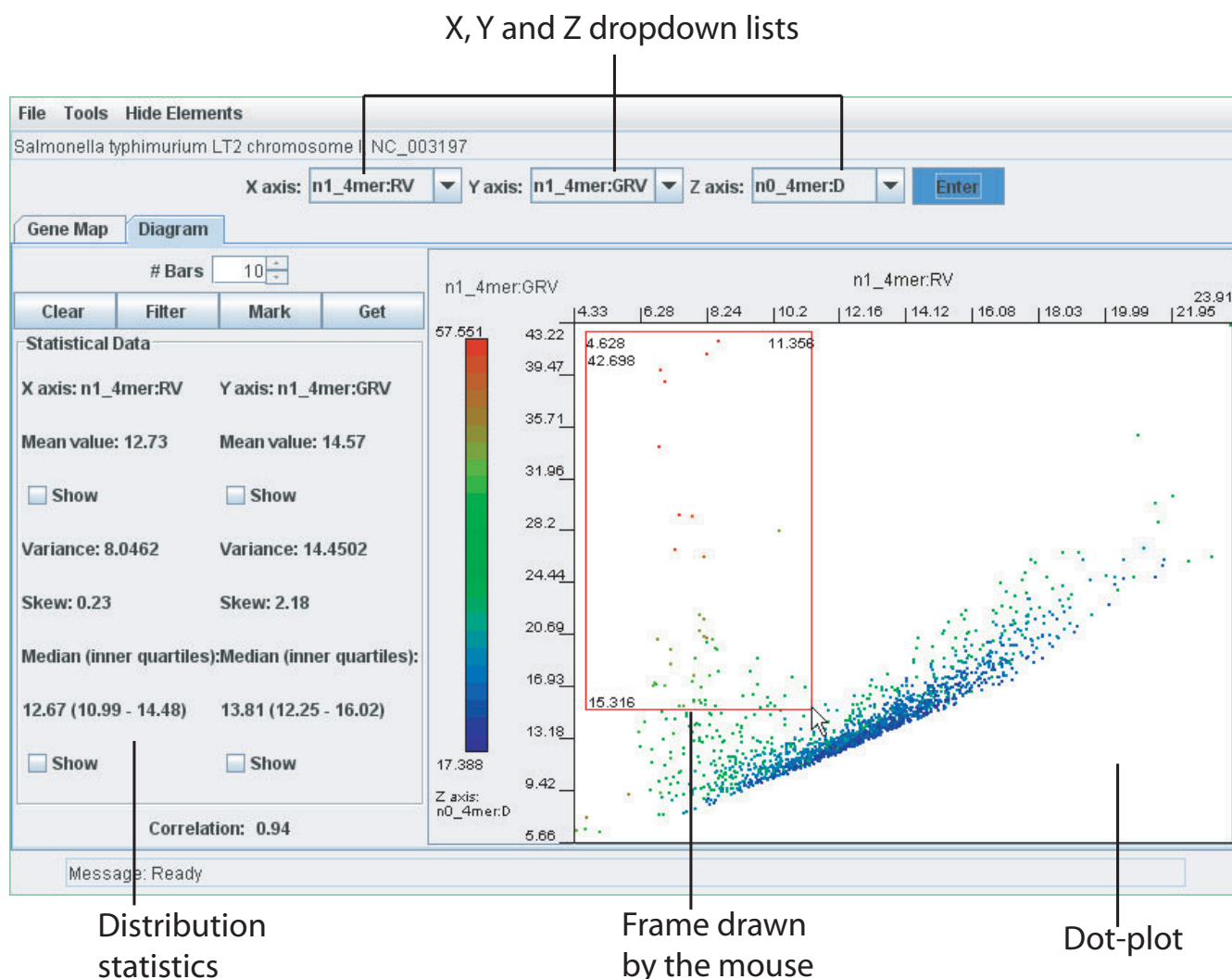
parameter. One may select an area on the colour bar to exclude all dots from the plot lying outside of the selected colour range (see an example in help files on-line). The hidden dots will not be selected by the 'Get' button. A more sophisticated way to filter genomic regions is provided by the 'Filter' button. An example will be discussed below.

The 'Mark' button enables genomic fragments to be selected by their coordinates and highlighted on the dot-plot. Click the 'Mark' button to open a dialog and enter

coordinates of one or multiple fragments (Fig. 4). Coordinates of each fragment must be added to the list by clicking the 'Add' button. Close the dialog by clicking 'OK'. The corresponding dots on the dot-plot will be highlighted as shown in Fig. 4.

Identification of divergent genomic islands

Several routines have been developed to identify the horizontally transferred genomic islands, genes for ribosomal RNA and proteins, non-functional pseudogenes and genes of other functional categories. All these routines are

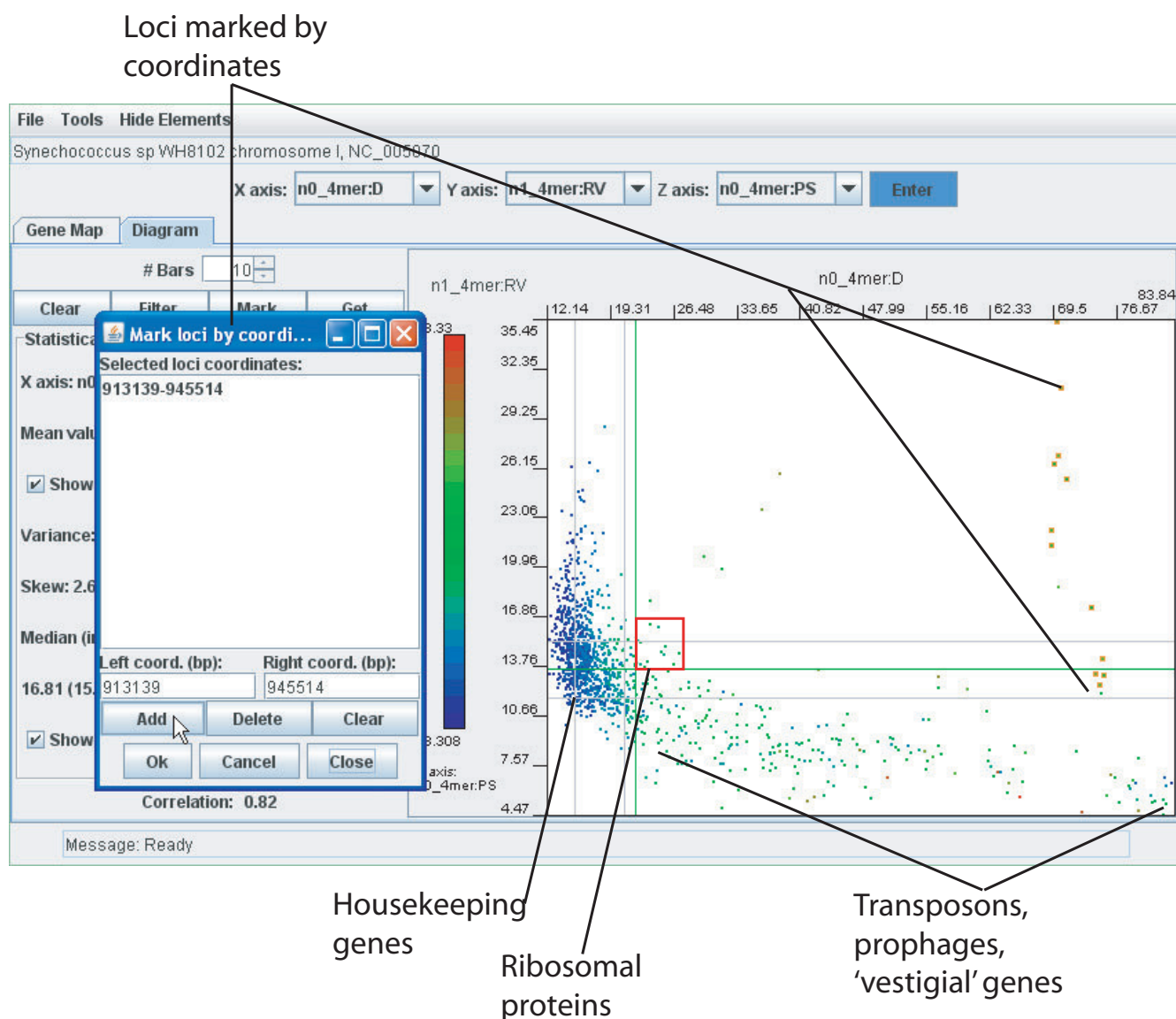
**Figure 3**

The 'Diagram' view. To draw a diagram, first select corresponding OU parameters using the dropdown lists and click the 'Enter' button. In this example n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively. Every dot on the dot-plot corresponds to a genomic fragment selected by the sliding window. Dots are spread and coloured in accordance with their values of the selected statistical OU parameters. Information for each dot may be found by one of the following methods: i) information for a dot under the mouse pointed by the mouse is shown in the 'Message' bar; ii) double clicking a dot returns us to the 'Gene map' tab with the corresponding genomic fragment highlighted; iii) framing the dots and clicking the 'Get' button opens a new applet window with the information about all selected regions. In this example the genomic regions of *Salmonella typhimurium* LT2 (NC_003197) that correspond to horizontally transferred genetic elements were selected (see discussion in the text).

described in detail with illustrations in supplementary web-pages (use the 'Help' link in the applet window).

The approach to identify inserts of foreign genomic elements by OU statistical parameters has been described recently [21]. While several algorithms allow identification of horizontally transferred genomic islands [8-13], the multiple oligomer parameters used in the SWGB even allows tentative attribution of genomic fragments (and,

given the right scale, genes or gene clusters) to different functional classes using only a FASTA sequence as input. However, the emphasis of the SWGB is not primarily its annotation capability, but its ability to display the natural internal variability of genome sequences. We use *Pseudomonas putida* KT2440, a known mosaic genome with 105 genomic islands above 4000 bp in length [26] as an example. Many of these features can be visualized at a glance using the SWGB without any in depth analysis (see

**Figure 4**

Identification of divergent genomic regions by plotting and highlighting. In this example the genome of *Synechococcus* sp. WH8102 was analysed. The parameters n0_4mer:D, n1_4mer:RV and n0_4mer:PS were selected for the X, Y and Z axes, respectively. The genomic regions covering the giant gene for the surface protein SwmB [29] were highlighted by entering the coordinates of this gene into the 'Mark loci by coordinates' dialog. The genomic regions enriched with i) housekeeping genes; ii) genes for ribosomal proteins; iii) vestigial genetic elements (comprising pseudogenes, transposons, prophages and IS-elements) are indicated.

Fig. 2). On the 'Diagram' view the parameters n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively, as we showed previously (see Fig. 3). Plotting local relative oligomer variance (RV) against global relative variance (GRV) basically shows the effect of normalization by global mononucleotide content. The core genome is then represented on the dot plot as the positive linear correlation line where $RV \approx GRV$ (Fig. 3). In other words, these fragments exhibit such

compositional closeness to the core genome that normalizing by local mononucleotide content does not have a different effect compared to normalizing by global content. These genomic fragments also exhibit a low distance from the genomic average; and are therefore coloured blue. Scattered dots lying peripheral to the expected strong linear correlation do not belong to the core genome and also have a higher distance from the genomic average and are hence coloured green. Using the filter set-

tings recommended in Fig. 5, twenty one fragments were found to be genomic islands (note that while border values of OU parameters are not the same for different genomes, the grading notches of the sliders represent relative values that allows identification of homologous regions in many different genomes). For a number of reasons, many more islands were found in a similar analysis by Weinelt *et al.* [26]. Firstly, the sliding window size of 8 kbp means many of the 4 kbp features from their analysis were not identified automatically. Furthermore, they were looking for all compositionally atypical regions, whereas here we restrict ourselves to horizontally transferred regions.

A known 40 kbp bacteriophage insertion [2586000–2626000] is, surprisingly, not among the genomic fragments selected in the SWGB using this filter. Although the prophage is still perceptible on the 'Gene Map' view (see a figure in the supplementary help web-pages), the OU parameters of the region do not differ markedly enough from the core sequence to be isolated automatically as a horizontally transferred region.

As the SWGB uses parameters that are based on comparison of local fragments to the global genomic average, strains with abundant insertions of homogenous DNA can confound this form of analysis. One example is the *Methanosarcina acetivorans* C2A genome which is composed of an estimated 25% of putatively horizontally acquired DNA, one of the highest amounts discovered to date [11]. As a result of these insertions, the genomic signature has been strongly influenced, resulting in a large amount of scatter and a poorly defined core genome on the plots. On the other hand, this type of analysis allows estimation of genome stability in a simple, multi parameter view (see the *Vibrio cholerae* N16961-O1-eltor example in the online help files). To conclude, filtering provides a convenient way to automatically isolate divergent genomic regions of interest. However, some regions may erroneously remain undetected due to possible amelioration of older inserts [27] or a higher level of noise in unstable genomes. However, many problematic genomic fragments can in some cases be easily attributed to functional gene categories using the SWGB 'Diagram' window (see Fig. 2).

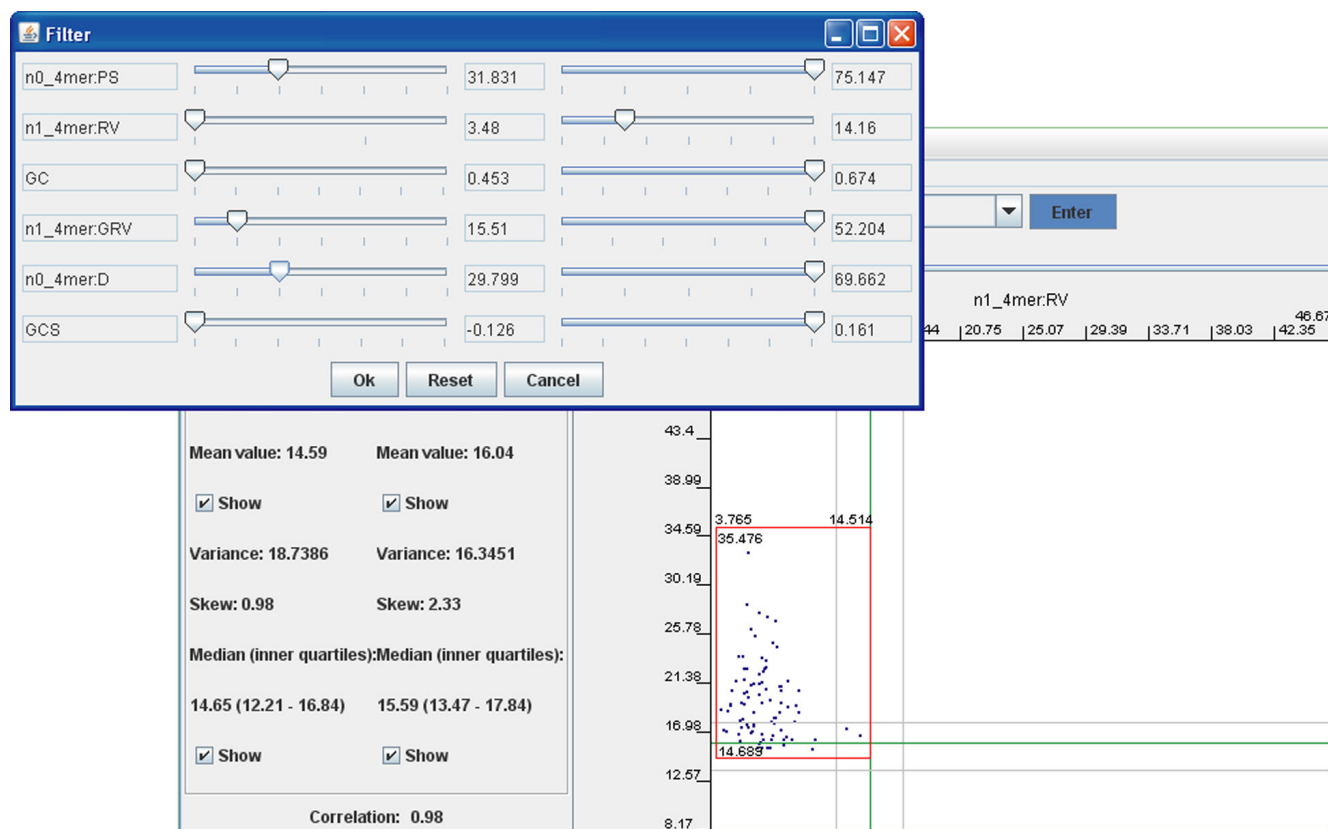


Figure 5

Filtering genomic regions by multiple parameters. Click the 'Filter' button to open a dialog as shown in the figure. Setting up border values of multiple OU statistical parameters allows more precise localisation of regions of interest.

Methodologies for discovering long modular genes have already been discussed in a previous publication [28]. Briefly, long genes display a particular tetranucleotide usage and can be discovered by plotting $n0_4mer:D$ (X axis) versus $n1_4mer:RV$ (Y axis). The positively linear correlated outlier fragments (towards the top right of the image) are often fragments of long genes with their characteristic repeats. An example using the gene encoding the 1.12 megadalton cell surface protein of *Synechococcus* sp. WH8102 [29] marked on the dot-plot is shown in Fig. 4. Ribosomal RNA operons (but not genes for ribosomal proteins) are characterized by extremely high pattern skew and a large distance from the core genome (Fig. 2). Thus, there is a tendency to find many genomic fragments containing rRNA genes coloured dark brown to red in the bottom right section of the 'Diagram' tab. The annotation for rRNA operons is not present in the database; therefore, these are seen in the 'Gene Map' tab as unannotated areas with high pattern skew (Fig. 2). Ribosomal proteins tend to be increasingly present at a slightly greater than average RV and above average D (see Fig. 2), which is in agreement with observations that highly expressed genes for ribosomal proteins have a highly specific codon usage compared to housekeeping genes of the organism [30]. The majority of genomic fragments form a cluster characterized by average and higher than average RV, stable OU patterns (low D) and low PS. These tend to be the core, or bulk genes and genomic regions with their typical tetranucleotide usage. Some other core sequence fragments spread from this area toward lower RV and less specific OU patterns (higher D and PS) – these are all characteristics of an unstable or randomly generated sequence [20]. These regions were found to be enriched with many hypothetical genes, prophages and transposons. (The data is not shown but is easily verified with any genome using the 'Get' button. Consider, for example, this area in the pseudogene rich *Mycobacterium leprae* TN or *Methanosaarcina acetivorans* C2A genomes [11,31], and the relatively homogenous *Alcanivorax borkumensis* SK2 genome [32].) These regions were thus categorized as rich in 'vestigial' genes in contrast to the core genome regions rich in housekeeping genes (Fig. 4).

It must be stressed that with an average length of genes being around 1 kbp and overlapping sliding windows of 8 kbp, one cannot expect precise separation of housekeeping and vestigial genes by the method described above. However, when analyzing an unknown DNA sequence prior to annotation, it may be helpful to identify genomic regions enriched with a higher proportion of these so called housekeeping genes and other regions rich in vestigial genes. These tentative results should be verified with other complementary algorithms such as BLAST, gene finding and annotation techniques.

The most important feature of the supplemented software available from the SWGB web-server for download is the ability to quickly and easily analyze a novel sequence on a local computer. The command-line Python program OligoWords is first used to analyse FASTA or GenBank formatted sequences. The program is available for download [33] in several packages as precompiled executable files and as Python source code. The command-line interface of the OligoWords program is shown in Fig. 6. Parameters such as oligomer length and window size can all be set depending on the sequence length and desired resolution (see Table 1 for suggestions). Since the SWGB is implemented as a Java applet, it can be run within a web browser locally. The HTML-embedded applet is available for download from the same FTP site [33] (select SeqWord_Viewer.zip). The output file from OligoWords is read into the SWGB via the 'Open' function of the 'File' menu, and the complete functionality of the online system is then available. For example, a new sequence can be analysed for ribosomal gene clusters, putative horizontally transferred elements or other regions of atypical DNA structure prior to the lengthy annotation step. A complete description of how to run the SWGB and OligoWords locally is presented in the online help files.

Conclusion

The SWGB applies novel OU statistics to visualize and discern divergent genomic regions. It has been extensively tested in practice for large scale genome analysis [32,34], and for identification and comparison of horizontally transferred genomic islands [35]. The applet is linked to a database of pre-calculated OU patterns of bacterial genomes (1243 complete sequences, including bacterial chromosomes, plasmids and some viruses were available at the time of manuscript submission, however, new sequences are regularly being added). The SWGB allows tentative annotation of the various divergent regions and provides overviews for use in comparative genomics. Users may download the command line version of the OligoWords program to analyze their own sequences. A packaged version of the SWGB allows users to view and manipulate their OligoWords results locally using a compatible web-browser.

Although there are several readily available tools for DNA compositional analysis, genomic island identification and large scale genome analysis [36-42], the SWGB surpasses previous approaches in making use of a wider range of parameters which allow identification of divergent genomic regions and even visual tentative attribution of these DNA fragments to various categories. We have found superimposition of these parameters to be more informative than a simple GC average or a relative OU frequency deviation since they allow discrimination of divergent genomic regions (large modular genes, ribosomal

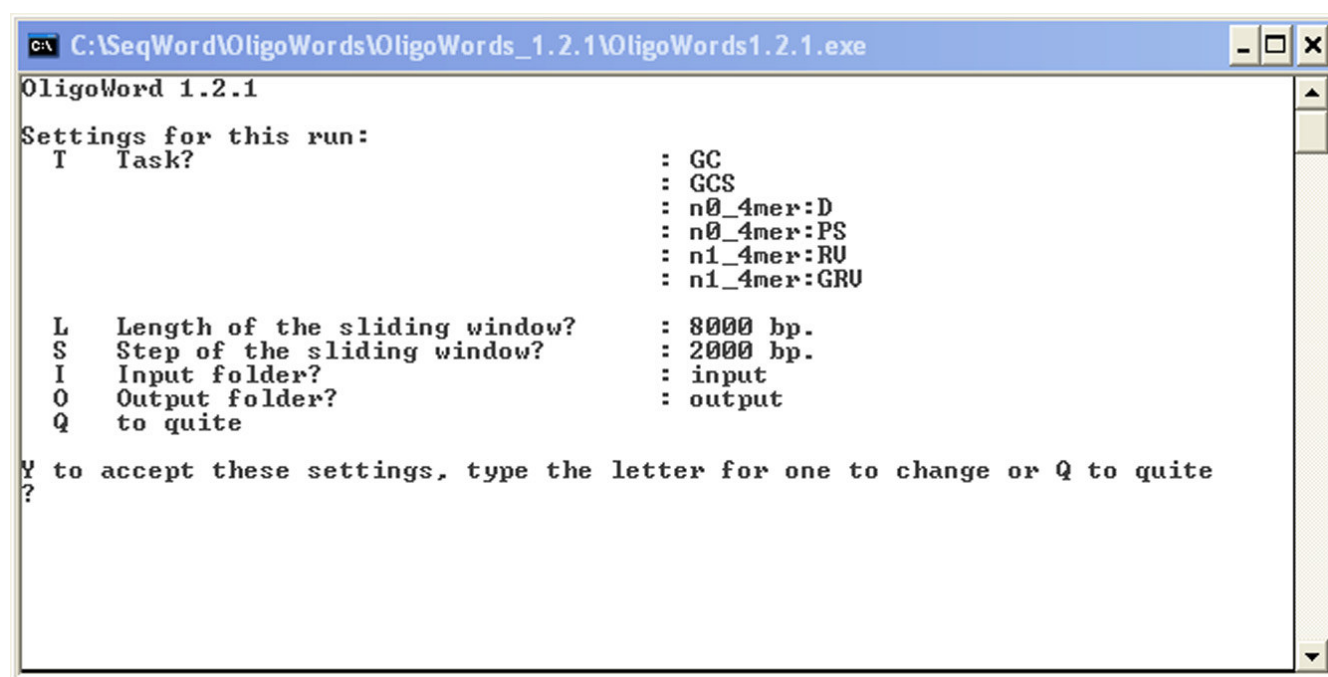


Figure 6

Command-line interface of the OligoWords program. To change the setting for the current run, type the option's letter and enter a new value as prompted. Users may change: T) the set of statistical OU parameters to be calculated for every local pattern; L) length of the sliding window; S) step of the sliding window; I) the name of the input folder that contains FASTA and/or GenBank files with source DNA sequences; and O) the name of the output folder where the result files will be stored.

RNAs, ribosomal protein clusters and the horizontally transferred genomic islands, see Fig. 2) all of which are characterized by an alternative OU composition relative to the core sequence. In addition, our approach provides some insight into the physicochemical state of the analysed DNA and the stability/state of flux of a genome as tetranucleotides exert a strong structural signal [20,21]. Consideration of flux inferred oligonucleotide usage is particularly interesting when comparing, for example, multiple replicating units of the same strain. Using the simple analysis described here, the second chromosome of *Vibrio cholerae* N16961-O1-eltor was demonstrated to be far less conserved than the first, with differences in mononucleotide content and distance from core genomic values implying a more heterogenous chromosome consistent with its role as a gene capture system [43].

Furthermore, no single oligonucleotide word size has been found to be optimal for all purposes, such as finding conserved or horizontally transferred DNA, plasmid host comparisons or testing distant homology [19]. The SWGB crucially provides the opportunity to analyse DNA sequences with various oligomer lengths and normalisation schemes. For example, genomic regions of particular interest may be multiply analysed with progressively

smaller oligomer sizes to provide more detailed information on oligomer usage in individual genes.

Availability and requirements

The SWGB applet is freely available to any researcher wishing to use it for non-commercial purposes via the Internet [22-24]. It has been tested on openSUSE 10.2, Gentoo Linux 2.6, Fedora Core 5 and Microsoft Windows XP workstations using Microsoft Internet Explorer 6.0, Maxthon 1.5.9, Mozilla Firefox 2.0, Mozilla SeaMonkey 1.1.1, Safari 3.0.4 for Mac, Konqueror 3.5.5 and 3.5.7, and Opera 9.10 browsers with Java 1.5. At the time of manuscript submission, a problem likely related to the local firewall was encountered with the Firefox browser on SUSE 10.2 and some other browsers (see 'Compatibility' link on the SWGB front page). The problem will be tackled in later releases of the SWGB. Feedback from users (addressed to the corresponding author) is very much appreciated.

Abbreviations

SWGB: SeqWord Genome Browser; OU: oligonucleotide usage; D: distance between two oligonucleotide usage patterns; PS: oligonucleotide usage pattern skew; RV: relative variance of the oligonucleotide usage; GRV: globally normalized relative variance of the oligonucleotide usage;

GC: guanine + cytosine content; GCS: guanine versus cytosine skew in DNA strands.

Authors' contributions

HG participated in development of the MySQL database and the SWGB web-site, Python programming, applet testing; ASR contributed by Java and PHP programming; CFD participated in development and support of the SWGB web-site, Help files and the applet testing; BT participated in development and testing of the OU statistical algorithms; ONR participated in development and testing of the OU statistical algorithms, supervision of the SeqWord project and Python programming.

Acknowledgements

This work was funded by the National Bioinformatics Network of South Africa <http://www.nbn.ac.za>. CFD is supported by the Deutsche Forschungsgemeinschaft grant DFG 653/3. The authors are thankful to Prof. Anton Nekrutenko from Penn State University (Pennsylvania, USA) for hosting the SeqWord applet at <http://seqword.bx.psu.edu/>.

References

- Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends Genet* 1995, **11**(7):283-290.
- Karlin S, Mrázek J, Campbell A: **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriol* 1997, **179**(12):3899-3913.
- Karlin S: **Global dinucleotide signatures and analysis of genomic heterogeneity.** *Curr Opin Microbiol* 1998, **1**:598-610.
- Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B: **Genomic signature: characterization and classification of species assessed by chaos game representation of sequences.** *Mol Biol Evol* 1999, **16**(10):1391-1399.
- Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ: **Evolutionary implications of microbial genome tetranucleotide frequency biases.** *Genome Res* 2003, **13**:145-155.
- Coenye T, Vandamme P: **Use of the genomic signatures in bacterial classification and identification.** *System Appl Microbiol* 2004, **27**:175-185.
- van Passel MW, Bart A, Luyf AC, van Kampen AH, Ende A van der: **The reach of the genome signature in prokaryotes.** *BMC Evol Biol* 2006, **6**:84.
- Mrázek J, Karlin S: **Detecting alien genes in bacterial genomes.** *Ann NY Acad Sci* 1999, **870**:314-329.
- Azad RK, Lawrence JG: **Use of artificial genomes in assessing methods for atypical gene detection.** *PLoS Comput Biol* 2005, **1**:e56.
- Becq J, Gutierrez MC, Rosas-Magallanes V, Rauzier J, Gicquel B, Neyrolles O, Deschavanne P: **Contribution of horizontally acquired genomic islands to the evolution of tubercle bacilli.** *Mol Biol Evol* 2008, **24**:1861-1871.
- Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P: **Detection and characterization of horizontal transfers in prokaryotes using genomic signature.** *Nucleic Acids Res* 2005, **33**:e6.
- Nakamura Y, Itoh T, Matsuda H, Gojohori T: **Biased biological functions of horizontally transferred genes in prokaryotic genomes.** *Nat Genet* 2004, **36**:760-766.
- Pride DT, Blaser MJ: **Identification of horizontally acquired elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis.** *Genome Let* 2002, **1**:2-15.
- Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T: **Informatics for unveiling hidden genome signatures.** *Genome Res* 2003, **13**:693-702.
- Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO: **Application of tetranucleotide frequencies for the assignment of genomic fragments.** *Environ Microbiol* 2004, **6**:938-947.
- Jernigan RW, Baran RH: **Pervasive properties of the genomic signature.** *BMC Genomics* 2002, **3**:23.
- Koski LB, Morton RA, Golding GB: **Codon bias and base composition are poor indicators of horizontally transferred genes.** *Mol Biol Evol* 2001, **18**:404-412.
- Wang B: **Limitations of compositional approach to identify horizontally transferred genes.** *J Mol Evol* 2001, **53**:244-250.
- Bohlin J, Skjerve E, Ussery D: **Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes.** *BMC Genomics* 2008, **9**:104.
- Reva ON, Tümmmler B: **Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns.** *BMC Bioinformatics* 2004, **5**:90.
- Reva ON, Tümmmler B: **Differentiation of regions with atypical oligonucleotide composition in bacterial genomes.** *BMC Bioinformatics* 2005, **6**:251.
- SWGB mirror site at the University of Pretoria in South Africa** [<http://www.bi.up.ac.za/SeqWord/mhhapplet.php>]
- SWGB mirror site at Hannover Medical School in Germany** [<http://genomics1.mh-hannover.de/seqword/genomebrowser/mhhapplet.php>]
- SWGB mirror site at Penn State University in the USA** [<http://seqword.bx.psu.edu/mhhapplet.php>]
- Reva O, Tümmmler B: **Oligonucleotide usage signatures of the *Pseudomonas putida* KT2440 genome.** In *Pseudomonas: Genomics and Molecular Biology Volume Chapter 3*. Edited by: Cornelis P. Caister Academic Press; 2008:43-64.
- Weinel C, Nelson KE, Tümmmler B: **Global features of the *Pseudomonas putida* KT2440 genome sequence.** *Environ Microbiol* 2002, **4**:809-818.
- Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44**:383-397.
- Reva O, Tümmmler B: **Think big – giant genes in bacteria.** *Environ Microbiol* 2008, **10**:768-777.
- McCarren J, Brahamsha B: **Swmb, a 1.12-megadalton protein that is required for nonflagellar swimming motility in *Synechococcus*.** *J Bacteriol* 2007, **189**:1158-1162.
- Puigbò P, Romeu A, Garcia-Vallvé S: **HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection.** *Nucleic Acids Res* 2008, **36**:D524-D527.
- Lawrence JG, Hendrix RW, Casjens S: **Where are the pseudogenes in bacterial genomes?** *Trends Microbiol* 2001, **9**:535-540.
- Reva ON, Hallin PF, Willenbrock H, Sicheritz-Ponten T, Tümmmler B, Ussery DW: **Global features of the *Alcanivorax borkumensis* SK2 genome.** *Environ Microbiol* 2008, **10**:614-625.
- SWGB FTP site** [<ftp://milliways.bi.up.ac.za/SeqWord/GenomeBrowser/>]
- Chen XH, Koumoutsis A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I, Morgenstern B, Voss B, Hess WR, Reva O, Junge H, Voigt B, Jungblut PR, Vater J, Süßmuth R, Liesegang H, Strittmatter A, Gottschalk G, Borriss R: **Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42.** *Nat Biotechnol* 2007, **25**:1007-1014.
- Klockgether J, Reva O, Tümmmler B: **Spread of genomic islands between clinical and environmental isolates.** In *Prokaryotic Diversity: Mechanisms and Significance* Edited by: Logan NA, Lapping-Scott HM, Oyston PCF. Cambridge: Cambridge University Press; 2006:187-200.
- TETRA** [http://www.megx.net/tetra_new/index.html]
- GENESTYLE** [<http://genstyle.imes.jussieu.fr/>]
- δp-WEB and Compare_Islands** [<http://deltarho.amc.nl/cgi-bin/bin/index.cgi>]
- The Horizontal Gene Transfer Database (HGT-DB)** [<http://www.tinet.org/~debb/HGT/>]
- IslandPath** [<http://www.pathogenomics.sfu.ca/islandpath>]
- Artemis** [<http://www.sanger.ac.uk/Software/Artemis/>]
- Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R: **Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models.** *BMC Bioinformatics* 2006, **7**:142.
- Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleischmann RD, Nierman WC, White O, Salzberg SL, Smith HO, Colwell RR, Mekalanos JJ, Venter JC, Fraser CM: **DNA sequence of both**

chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 2000, **406**:477-483.

44. **NCBI, Complete Microbial Genomes** [<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



3. Global distributions of oligomers across the prokaryotic kingdom

3.1 Background

Microbiologists have traditionally focussed their attentions on a gene, operon or pathway of interest. This gene-centric approach, which has clear relevance for function, has led to many of the discoveries the field has achieved to date. The same is true in the post-genomic era. Common algorithms such as Blast were chiefly designed and adapted for fast gene level searches. Widespread adaptation of this paradigm has, however, led most bioinformaticians and genomic scientists to concentrate on larger DNA size ranges. Due to the development of oligomer based hybridisation chips such as microarrays much effort has been devoted to assessing sensitivity and specificity in larger oligomers of perhaps 20-50 bp in size. This effort not only includes industry, as academics have also developed a wide range of custom chips for various purposes. Uses include DNA hybridisation of rRNA genes in the human intestinal tract (Rajilic-Stojanovic et al. 2009), and microarrays for gene expression analysis (Dufva 2009). Despite all this interest in their "big brothers", only a few groups have thoroughly investigated short to medium sized oligomers and possible *in silico* applications (Karlin et al. 1997, Weinelt et al. 2002, Pride et al. 2003, van Passel et al. 2005, Mrázek 2009, Willner et al. 2009). To our knowledge, no researchers have investigated either single medium sized oligomers commonly overrepresented in many bacteria, or distributions of overrepresented oligomers across all sequenced genomes. Approximately 700 genomes were available at the time of this thesis, which, though heavily biased towards human pathogens, cover a broad range of GC content and were sufficient to obtain some general rules (Koonin and Wolf 2008). We describe here the factors involved in selecting oligonucleotides to be included in the analysis and apply these to all sequenced prokaryotes via two different approaches.

3.1.1 Measuring overrepresentation

One critical point when analysing short oligomers is overrepresentation. Any short oligomer is likely to be present in a sequence purely by chance. Longer oligomers such as 8mers are also predicted to be present in low copy numbers in most sequences by combinatorics, while 14mers are not. This approach assumes all oligomers are equally and randomly distributed, but would require that genomes are extremely homogenous and contain equal proportions of each nucleotide. Zero order Markov models such as those used in OligoCounter (part 1) are a step up in complexity over those assuming random distributions, and control for GC content in the oligomer assessed as well as

that of the whole genome (Almagor 1983, Pride et al. 2003, Bohlin et al. 2008). No one predictive computational model of DNA sequence is appropriate and superior for all situations (Mrázek 2009), and more complicated models may also obscure results by excessive normalisation for biases such as dinucleotide signal. For example, Pride et al. (2003) found the relatively simple zero order Markov models preserved more phylogenetic signal than complicated Markov chain models in a large scale phylogenetic study. Two different approaches were used in the following manuscripts, a random model based purely upon copy numbers normalised by megabase, and an exhaustive statistical approach based on most oligomers present in a genome.

3.2 About the manuscripts

The first manuscript, entitled "Abundant oligomers common to most bacteria", was inspired by a serendipitous finding when analysing oligonucleotides in the genus *Pseudomonas* for the paper presented in Part 1. While assessing metagenomic specificity, oligomers that are highly overrepresented in all *Pseudomonas* sp. were discovered. Thus we were curious to investigate whether oligomers common to all, or most prokaryotes could be found. To do so, we hypothesised that if overrepresented oligomers were found in various highly divergent prokaryotic lineages they may be common to all bacteria. The main finding was 12 oligomers, which could be condensed down to one almost complementary oligomer. The four most common homopurine variants 5'-RAAGAARA contained zero, one or two guanine nucleosides at positions one and seven, while the eight most common homopyrimidines 5'-YTTYTTYT harboured zero, one or two cytosine nucleosides at positions one, four and seven. These 12 octanucleotides occurred six- to thirtyfold more frequently in our dataset of 684 chromosomes than expected for a randomly selected octanucleotide. After analysing the distribution of oligomers in the genomes, strikingly none were found to aggregate in any one coding frame. A potential role in DNA structure was thus implied, and structural parameters were found to be extremely rigid. Interestingly, short poly-A stretches with a G in a central position have been found to form triple helical structures, even at common physiological temperatures (Rajagopal and Feigon 1989). Given that these oligomers are overrepresented in almost all bacterial lineages, their extreme structures and their polypurine composition which is prone to unusual helix formation, triple helices may well be more common in prokaryotic genomes than previously thought.

The second manuscript, "Oligonucleotide signatures of the bacterial world", discusses oligomer distributions across all 22 phylogenetic lineages represented by sequenced genomes at the

time. Standard overrepresented data restricted by chi-squared statistics from OligoCounter were not used but threshold criteria were set very low, then all data using were reductively scanned using large databases for analysis. Objectives were analysis of 8-14 bp suitability for use in a metagenome analysis system, and at the same time genomic analyses of all sequenced prokaryotes. This extensive approach was quite time-consuming, yet brought a number of surprises. The main result was that two distinct classes of signature words could be differentiated, biased towards coding and non-coding regions respectively. Of the coding regions, large repeat regions were identified across many genomes. It was suggested that these could be biochemically neutral blocks of amino acids. The greatest numbers of overrepresented words were found in the betaproteobacteria and actinobacteria, while chlamydiae and crenarchaeotes carried the least. Statistical methodologies were defined for assessing the variances and clustering of the oligomers. The manuscript then tentatively discusses the use of the identified oligomers in metagenomics.

The first manuscript, "Abundant oligomers common to most bacteria", is in revision at Plos One and the second, "Oligonucleotide signatures of the bacterial world", is in revision at Environmental Microbiology at the time of writing. My contribution to the first as lead author included data collection, all analyses, and programming and sections of the manuscript. Burkhard Tümmeler wrote most parts of the manuscript. In the second manuscript I was more loosely involved, contributing raw data, some supervision, and reviewing the document.

3.3 References

- Almagor, H. (1983) A Markov analysis of DNA sequences. *J Theor Biol* **104**(4) 633-645.
- Bohlin, J., Skjerve, E. & Ussery, D. W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* **4**(4) e1000057.
- Dufva, M. (2009) Introduction to microarray technology. *Methods Mol Biol* **529**: 1-22.
- Karlin, S., Mrázek, J. & Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12) 3899-3913.
- Koonin, E. V. & Wolf, Y. I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* **36**(21) 6688-6719.
- Mrázek, J. (2009) Phylogenetic signals in DNA composition: limitations and prospects. *Mol Biol Evol* **26**(5) 1163-1169.

- van Passel, M. W. J., Luyf, A. C. M., van Kampen, A. H. C., Bart, A. & van der Ende, A. (2005) Deltarho-web, an online tool to assess composition similarity of individual nucleic acid sequences. *Bioinformatics* **21**(13) 3053-3055.
- Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2) 145-158.
- Rajagopal, P. & Feigon, J. (1989) Triple-strand formation in the homopurine:homopyrimidine DNA oligonucleotides d(G-A)₄ and d(T-C)₄. *Nature* **339**(6226) 637-640.
- Rajilić-Stojanović, M., Heilig, H. G. H. J., Molenaar, D., Kajander, K., Surakka, A., Smidt, H. & de Vos, W. M. (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ Microbiol.* **11**(7) 1736-1751.
- Weinel, C., Ussery, D. W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C. & Tümmler, B. (2002) Comparative Genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: Orthologs, Codon Usage, Repetitive Extragenic Palindromic Elements, and Oligonucleotide Motif Signatures. *Genome Letters* **4**, 175-187.
- Willner, D., Thurber, R. V. & Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol.* **11**(7) 1752-1766.

Abundant oligonucleotides common to most bacteria

Colin F. Davenport and Burkhard Tümmler*

Clinical Research Group, OE6711, Hanover Medical School, Carl-Neuberg-Strasse 1,
D-30625 Hanover, Germany.

*Corresponding author. Mailing address: Clinical Research Group, OE6711, Hanover
Medical School, Carl-Neuberg-Strasse 1, D-30625 Hanover, Germany. Phone: 0049 511
5322920, Fax 0049 511 5326723. E-mail: tuemmler.burkhard@mh-hannover.de

Running title: Abundant oligonucleotides

Keywords: triple helix, abundant oligonucleotide, homopurine, evolutionary conservation

Abstract

Background

Bacteria show a bias in their genomic oligonucleotide composition far beyond that dictated by G+C content. Patterns of over- and underrepresented oligonucleotides carry a phylogenetic signal and are thus diagnostic for individual species. However, global distributions of oligomers have received little attention to date.

Principal Findings

Here we report a presumably ancient set of oligomers conserved and overrepresented in nearly all branches of prokaryotic life, including Archaea. These oligomers can be condensed to a single homopurine octamer 5'-RAAGAARA and its complement, with variations. They do not aggregate in any coding frame, implying a role in DNA structure. Structural parameters indicate these oligonucleotides to be an extreme and rigid form of B-DNA prone to forming triple stranded helices under common physiological conditions.

Conclusion

Homopurine and homopyrimidine oligomers exhibit distinct and unusual structural features and are present at high copy number in nearly all phylogenetic groups. This fact suggests a non-neutral role of these oligonucleotides for bacterial genome organization that has been maintained throughout evolution.

Introduction

Bacterial genomes may vary widely in nucleotide content. This is most readily observable in region specific G+C content [1]. However, higher order oligonucleotide composition fluctuates far more within a genome than simple G+C content would suggest [2,3]. This composition may be maintained due to replication and repair machinery, restriction modification or DNA structural restrictions [4]. As such, oligonucleotide biases represent an additional source of information which can be used to characterise a genome. For example, patterns of over- and underrepresented oligonucleotides carry a phylogenetic signal and are thus diagnostic for many individual species [2,3,5]. Oligonucleotide signatures have also been used in a variety of other areas, such as phylogenetic classification of metagenomic sequences [6]. However, global distributions of intermediate length oligomers have escaped detailed investigation to date. We here analyse oligonucleotides overrepresented across a large set of sequenced genomes from diverse lineages of the bacterial world.

Materials and Methods

The program OligoCounter [3], (available at <http://webhost1.mh-hannover.de/davenport/oligocounter/>), was used to count overrepresented 8-14 bp oligomers in the whole genomes of 684 chromosomes available from the NCBI FTP site (February 2008).

To be considered overrepresented, OligoCounter thresholds were set to retain oligomers present at least 31 times in the genome with a χ^2 -value of 100 or more. We estimate 15.2 ($= 10^6 / 4^8$) copies of a random octamer are expected to be present in each Mb of a prokaryotic genome. Thus, a threshold of twice this value (31) together with the χ^2 cut-off

restricted the oligomers we analysed. It should however be noted that median values of the located oligomers were far in excess of this value (Figure S1). χ^2 -statistics were calculated according to the following formula [7]:

$$\chi^2 = \frac{(C_{obs} - C_{exp})^2}{C_{exp}}$$

where C_{obs} is the observed count of words and C_{exp} is the expected count of words.

Expected counts of oligomers were derived by a zero order Markov model [8]:

$$E = N \times A^a \times C^c \times G^g \times T^t$$

where N is the genome size in nucleotides, A is the proportion of adenine in the genome and a is the number of adenines in the oligo, and so on for the other bases. The χ^2 - statistic is here not used as an indicator for statistical significance but merely of level of overrepresentation of each oligomer, otherwise Bonferroni corrections for multiple tests would have been necessary. A random 6 Mb genome with 50% G+C content was generated as a control and demonstrated to have no oligonucleotide biases at a χ^2 -value of 100.

Oligomer selection strategy

We compared all oligomers from five strains belonging to a selection of the most phylogenetically distinct lineages (Plantomycetes, Spirochaetes, Chlamydiae, Bacteroidetes/Chlorobi and Cyanobacteria). Our hypothesis was that oligomers common to all of these “outlier” taxa would also be found in many others. This process was repeated with two sets of strains of differing G+C content to confirm the robustness of this hypothesis:

Set 1 (Average G+C = 42.4%): *Rhodopirellula baltica* SH 1, *Leptospira interrogans* serovar Copenhageni str. Fiocruz L1-130, *Chlamydia trachomatis* A/HAR-13, *Bacteroides fragilis* NCTC 9343, *Prochlorococcus marinus* MIT 9211.

Set 2 (Average G+C = 51.2%): *Rhodopirellula baltica* SH 1, *Treponema pallidum* Nichols, *Chlamydophila pneumoniae* AR39, *Porphyromonas gingivalis* W83, *Gloeobacter violaceus* PCC7421.

Both sets included *Rhodopirellula* as this was the only sequenced strain in the Plantomycetes at the time. Set 1 resulted in 73 while set 2 produced 31 oligomers overrepresented in all five strains. Oligonucleotides were then aligned using the ClustalW multiple sequence alignment algorithm [9] in Jalview [10]. Homopurine and homopyrimidine sequences were separated because they were obviously distinct from one another, with 5'-AAGAA or 5'-TTCTT cores generated as complementary consensus sequences. Including all of the observed variable alignment positions immediately around these cores resulted in the analysed 12 oligomer sequences (Table 1). As an example for completion of the AAGAA core, the multiple sequence alignments displayed possible guanine or adenine at position 1, guanine at position 7 and only adenines at position 8. Thus four oligomers AAAGAAAA, AAAGAAGA, GAAGAAGA and GAAGAAAA were scanned for in the dataset of all sequenced genomes, thus covering all of these observed possibilities.

Statistical analysis of coding regions

Percentages of genomes from the February 2008 NCBI RefSeq genome collection which are coding were calculated with an in-house script (available from the authors). This resulted in a simple percentage for each genome. The percentage of abundant oligomers also occurring in coding regions was then calculated. Wilcoxon rank sum tests were used to compare coding distributions in all NCBI genomes (n=684) against the percentage of abundant oligomers occurring in coding regions. Tests were carried out and distributions plotted using the statistical environment R [11]. A further script calculated the coding frame

which each oligomer within an ORF was present in using genome position and annotation information.

Results and Discussion

Sequenced bacterial and archaeal genomes were scanned for globally overrepresented 8- to 14-mers, a previously largely ignored class of oligomers. The search revealed a single set of nearly complementary homopurine and homopyrimidine octanucleotides as the statistically most overrepresented widespread oligomers (Figure 1; Supplementary Information, Figures S1, S2; Tables S1, S2, S3, S4). The four most common homopurines 5'-RAAGAARA harbour zero, one or two guanine nucleosides at positions one and seven, while the eight most common homopyrimidines 5'-YTTYTTYT carry zero, one or two cytosine nucleosides at positions one, four and seven (Table 1). These 12 octanucleotides occur six- to thirtyfold more frequently in a dataset of 684 chromosomes than expected for a randomly selected octanucleotide (Figure S1).

Individual prokaryotic genomes harbour between a few dozen to up to several thousand copies of each of the 12 octanucleotides in one megabase of sequence (Figure S1). The complementary pair 5'-AAAGAAAA and 5'-TTTCTTTT were the two most abundant and widespread octanucleotides that each occurred at a frequency of 0.01% or more in about half of analysed bacterial and archaeal chromosomes (Figure S1). The abundant octanucleotides were randomly distributed between the three reading frames (Figure S3) implicating that neither codon usage bias nor highly common tripeptides [3] account for the high frequency of the oligomers.

The distribution of these oligos were compared between different replicons of a single genome for the eight relevant genera. In most genera these oligomers are not conserved in both replicons (*Burkholderia*, *Ralstonia*, *Brucella*, *Deinococcus*, *Paracoccus* and *Rhodobacter*). Two genera display exceptions: three of the four *Vibrio* species contain the four overrepresented RAAGAARA oligomers in both chromosomes, and four of four *Leptospira* strains contain overrepresented RAAGAARA oligomers in both replicons. While compositional differences between replichores [12] are undoubtedly interesting, the variability of these data precludes any further discussion.

Oligonucleotide usage is not neutral because of structural, functional, biological and coding constraints [2,4,5,13]. Since the abundant octanucleotides show no preference for any position in the reading frame (Figure S3), their abundance most likely does not reflect any transcriptional or translational demands, but rather results from structural and functional features of the oligomer within the chromosome.

Structural properties of oligonucleotides have been predicted from computational models [14,15] that are based on the crystal structures of double helical DNA oligomers [16], experimental studies on DNA flexibility and bending [17], empirical energy functions [18], and quantum mechanical calculations [19]. The structural scales derived from these empirical or theoretical models assign high propeller twist (-20°) and low flexibility ($0.3 \text{ \AA}$) to the AA/TT dinucleotide steps in the 12 abundant A-rich homopurines and T-rich pyrimidines found in our analysis, whereas the AG/CT steps should show intermediate conformational flexibility (-13° , $0.7 \text{ \AA}$) [20,21]. Furthermore, the conformation of the individual A-T base-pairs in each oligomer duplex are predicted to be largely independent of the sequence context, because the conformational properties of all possible neighbouring steps are known to be compatible [15].

Another feature of homopurines and homopyrimidines is their ability to form a triple-stranded helix [22-26]. One homopyrimidine tract forms conventional Watson-Crick base-pairs with the homopurine tract and the second homopyrimidine strand is Hoogsteen base-paired in the major groove to the homopurine strand. Two complementary homopurine-homopyrimidine octanucleotides are sufficient to induce this phenomenon [24]. Triple helix formation is known to be disfavoured in pure oligo-dA tracts, but the insertion of a single central guanine nucleoside has been shown to lead to observable triplex formation at neutral pH [26]. This sequence motif of a single G flanked by oligo(dA) is identical with the consensus sequence of widespread overrepresented oligomers found in this analysis of prokaryotic genomes (see Figure 1).

Aside from the discovery of *H*-DNA, a triple-stranded structure that forms under superhelical stress in plasmids [27], the biological role of triple helix formation has remained elusive. The finding that a single class of nearly complementary homopurine and homopyrimidine octanucleotides represents the statistically most overrepresented oligomer in prokaryotic genomes suggests that triple helix conformations may occur more frequently in nature than hitherto assumed. Short stretches of triplex DNA or hybrids of RNA with duplex DNA may exist in numerous archaeal and bacterial chromosomes, at least in mesophilic and psychrophilic microorganisms growing at lower temperatures where triplexes are more stable.

The 12 oligonucleotides were found to be overrepresented in chromosomes of nearly all phylogenetic groups (see the pie charts in Figure S2). A few taxonomic exceptions were noted. Most actinobacteria, deinococci and many metabolically versatile G+C-rich β - and γ -proteobacteria carried all 12 oligomers at frequencies below 31 per Mb (Table S1), indicating that stretches of purines and pyrimidines have not been positively selected in these clades.

In summary, the most widespread statistically overrepresented oligonucleotides in archaea and bacteria belong to a single class of A-rich homopurines and T-rich

homopyrimidines that are characterised by low conformational flexibility and the ability to form triple helices. Conservation of these oligomers in diverse taxonomic lineages implies an early origin and possible non-neutral role.

Acknowledgements

C.D. is a member of the International Research Training Group 'Pseudomonas Pathogenicity and Biotechnology' sponsored by the Deutsche Forschungsgemeinschaft. We would like to thank Lutz Wiehlmann for assistance with the analysis.

References

1. Dobrindt, U, Hochhut, B, Hentschel, U, Hacker, J (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2: 414-424.
2. Karlin S (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* 1: 598-610.
3. Davenport CF, Wiehlmann L, Reva ON, Tümmler B (2009) Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ Microbiol* 11: 1092-1104.
4. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 13: 145-158.
5. Reva ON, Tümmler B (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* 5: 90.
6. Mrázek J (2009) Phylogenetic Signals in DNA Composition: Limitations and Prospects *Mol Biol Evol* 26: 1163-1169.
7. Greenwood PE, Nikulin MS (1996) *A Guide to Chi-Squared Testing*. Wiley, New York.
8. Almagor H (1983) A Markov analysis of DNA sequences. *J Theor Biol* 104: 633-645.
9. Larkin MA, Blackshields G, Brown NP et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

10. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton, GJ (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189-1191.
11. R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
12. van Passel M, Bart A, Luyf A, van Kampen A, van der Ende A (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics*, 7: 26.
13. Bohlin J, Skjerve E, Ussery DW (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* 4: e1000057.
14. Baldi P, Baisnée PF (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* 16: 865-89.
15. Packer MJ, Dauncey MP, Hunter CA (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J Mol Biol* 295: 85-103.
16. El Hassan MA, Calladine CR (1997) Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Phil Trans Roy Soc Ser A* 355: 43-100.
17. Brukner I, Sánchez R, Suck D, Pongor S (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* 14: 1812-1818.
18. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 95: 11163-11168.

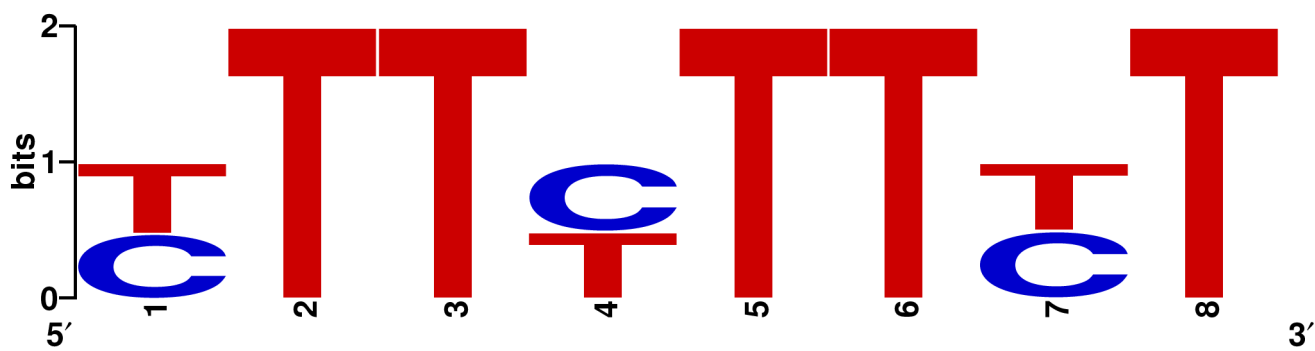
19. Ornstein RL, Rein R, Breen DL, MacElroy RD (1978) An optimised potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers* 17: 2341-2360.
20. El Hassan MA, Calladine CR (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *J Mol Biol* 251:648-664.
21. El Hassan MA, Calladine CR (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol* 259:95-103.
22. Duca M, Vekhoff P, Oussedik K, Halby L, Arimondo PB (2008) The triple helix: 50 years later, the outcome. *Nucleic Acids Res* 36: 5123-5138.
23. Felsenfeld G, Rich A (1957) Studies on the formation of two- and three-stranded polyribonucleotides. *Biochim Biophys Acta* 26: 457-468.
24. Rajagopal P, Feigon J (1989) Triple-strand formation in the homopurine:homopyrimidine DNA oligonucleotides d(G-A)₄ and d(T-C)₄. *Nature* 339: 637-640.
25. Roberts RW, Crothers DM (1996) Prediction of the stability of DNA triplexes. *Proc Natl Acad Sci U S A* 93: 4320-4325.
26. Sandström K, Wärmländer S, Gräslund A, Leijon M (2002) A-tract DNA disfavors triplex formation. *J Mol Biol* 315: 737-748.
27. Mirkin SM, Lyamichev VI, Drushlyak KN, Dobrynin VN, Filippov SA, Frank-Kamenetskii MD (1987) DNA H form requires a homopurine-homopyrimidine mirror repeat. *Nature* 330: 495-497.
28. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14: 1188-1190.

Table 1. The most widespread overrepresented oligomers and data on their overrepresentation.

Widespread oligomer	Number of chromosomes oligomer overrepresented in (n=684)	Median copy number per Mb*
AAAGAAAA	374 (55%)	166 (88 - 241)
AAAGAAGA	293 (43%)	111 (77 - 143)
CTTCTTCT	346 (51%)	68 (51 - 89)
CTTCTTTT	290 (42%)	120 (89 - 159)
CTTTTCT	357 (52%)	107 (70 - 138)
CTTTTTTT	243 (36%)	136 (87 - 197)
GAAGAAAA	372 (54%)	95 (56 - 134)
GAAGAAGA	463 (68%)	62 (46 - 82)
TTTCTTCT	272 (40%)	106 (82 - 144)
TTTCTTTT	354 (52%)	167 (98 - 244)
TTTTTCT	340 (50%)	144 (78 - 206)
TTTTTTTT	129 (19%)	143 (59 - 249)

*Quartiles are quoted in brackets

Figure 1. Sequence logos of the overrepresented oligomers. Figures were created with Weblogo [28] using the median number of chromosomes oligomers are overrepresented in (n for the purine group = 4 oligomers with a total of 1784 summed median copies, n for the pyrimidine group = 8 oligomers with 3097 summed median copies. Summed medians were derived from the median copy number of each oligomer across all genomes in which it was overrepresented, which were then added for each oligomer to create a weighted sequence logo).



Oligonucleotide signatures of the bacterial world

Warren Emmett¹, Colin F. Davenport^{1,2}, Phillip Labuschagne¹ and Oleg N. Reva^{1*}.

¹Bioinformatics and Computational Biology Unit, Department of Biochemistry,
University of Pretoria, Hillcrest, Pretoria 0002, Republic of South Africa.

²Klinische Forschergruppe, OE 6711, Medizinische Hochschule Hannover, Carl-
Neuberg-Strasse 1, D-30625 Hanover, Germany.

*For correspondence: Dr. Oleg N. Reva, tel: +27-12-420-5810; fax: +27-12-420-5800; e-mail: oleg.reva@up.ac.za

Running title: bacterial oligonucleotide signatures

Summary

For the first time the distribution of overrepresented oligonucleotides of lengths 8 to 14 base pairs was investigated on a genome-wide scale in 438 completely sequenced bacterial chromosomes. Two types of signature words were identified, which show a biased distribution in coding and non-coding genomic regions respectively. Although the biological function of these repeats remains unclear, it is apparent that these global repeats are more abundant in bacteria characterized by a more versatile metabolism and larger genome size. The coding signature repeats seem to serve as biologically neutral ‘bricks’ to fill gaps between the structural core and active sites of bacterial proteins. A genome linguistic analysis was performed to investigate whether these global repeats

were specific enough to distinguish DNA reads from different bacterial species within a metagenomic dataset. Several algorithms for the selection of discriminative words were proposed and applied to a structured database of overrepresented oligonucleotides (words) in order to determine candidates that can be used as genomic signatures. Based on the fact that at least one oligonucleotide from the template of selected signature words is expected to occur within every 10 bp of genomic sequence this approach appears promising for the processing of metagenomic datasets. This work focuses on the analysis of specificity for globally distributed oligomers and computational tools for processing real datasets will be made available in the future.

Introduction

Signature oligonucleotides are defined in this paper as oligonucleotides of intermediate lengths (between 8 and 14 base pairs) occurring frequently and evenly distributed throughout bacterial genomes. It has been shown that *Pseudomonas* species can be characterized with unique profiles of overrepresented words that may help to distinguish DNA fragments that originated from different genomes (Davenport *et al.*, 2009). Another interesting discovery was the existence of two distinct classes of abundant 8-14mer oligonucleotides with biased distribution in coding and non-coding genomic regions. Various short interspersed repeating oligonucleotides have been discovered in bacteria (Dimri *et al.*, 1992; Mancuso *et al.*, 2007). Of these the most well known are repetitive extragenic palindromic (REP) elements (Dimri *et al.*, 1992), the bacterial variable-number tandem repeats (Chang *et al.*, 2007; van Belkum 2007) and the Enterobacterial Repetitive Intergenic Consensus (ERIC) sequences (Hulton *et al.*, 1991). Furthermore, a

group of repeats referred to as regularly spaced repeats or CRISPR have been found widespread throughout different phylogenetic groups, being present in the majority of Archaea and in several members of the cyanobacteria and proteobacteria lineages (Mojica *et al.*, 2000; 2005). These examples demonstrate that several types of oligonucleotides are abundant in bacterial genomes and may be used to distinguish phylogeny. However, to our knowledge this paper reports the first large-scale study of the distribution of genome-wide overrepresented signature oligonucleotides, which may be used for identification of genomic fragments.

The investigation of bacteria and their environments have always been of great interest to the scientific community. Identification of the constituent bacterial species within environmental samples is a highly complex process as up to 99% of all bacteria in an ecological environment cannot be cultured in the laboratory (Mongodin *et al.*, 2005), whereas DNA is extracted and analyzed directly from the sample avoiding difficulties in culturing bacterial species. Metagenomic sequencing, however, leads to problems in determining the taxonomic origin of different reads. Current forms of sequence and species identification have proved inadequate within this context, as a large percentage of unique genomic signatures occur only within highly restricted regions and the loss of this segment would result in the loss of signal (Rodriguez-Valera 2002). By using oligonucleotides that occur uniformly throughout the genome, this study attempts to make identification of bacteria more robust against partial genome loss. Thus, short overrepresented oligonucleotides (2-6 bp in length) have been successfully used to identify genomic fragments (Karlin 1998; Teeling *et al.*, 2004; McHardy *et al.*, 2007; Nasser *et al.*, 2008). However, the short lengths of the oligomers may decrease specificity

of identification. Longer oligonucleotides (from 8 to 14 bp) are likely to be more specific. A major benefit of using overrepresented oligonucleotides is that it does not require the sequence to be assembled, which is a time consuming and error-prone process (DeLong 2005).

The aim of this work was to study the distribution of signature oligonucleotides in the bacterial world as a whole by linguistic analysis of completely sequenced bacterial genomes. We report on distributions and possible biological meaning of oligonucleotide repeats in coding and non-coding regions across 438 completely sequenced species and test the applicability of oligonucleotide signatures for species identification in environmental samples.

Results

Oligonucleotides extracted by the program OligoCounter (Davenport *et al.*, 2009) from 438 sequenced bacterial chromosomes were then used to populate the database. For each oligonucleotide the distribution score (DS) was calculated:

$$DS = \frac{100000}{\sqrt{\mu^2 + \sigma^2}} \quad (1)$$

where μ indicates the average number of bp (spans) between the repeated words in the sequence (i.e., $\mu = \text{sequence_length}/\text{number_of_words}$); and σ is the standard deviation of the span lengths. The DS for a word increases in value when there is a high frequency of occurrence (μ is minimal) and the words are evenly distributed (σ tends to 0). Thus multiple repeats of a word clustered in a specific genomic locus will be scored poorly as it characterizes the locus but not the genome as a whole.

Global distribution of signature words and definition of class templates

In this work we analyzed frequencies of oligonucleotides of intermediate lengths (between 8 and 14 bp). Every genome may be characterized by a specific set of signature words which were identified as words with DS higher than a set threshold. Statistics regarding the signature words selected for each genome at a DS threshold of 0.1 are shown in supplementary Table S1.

In order to compare species effectively, templates of signature words were constructed for bacterial phyla on different taxonomic levels. Thus, 22 templates were created for the phylogenetic classes as identified by GenBank (Benson *et al.*, 2007). The sets of signature words identified for all class members were combined into class templates. If an oligonucleotide was identified as a signature for at least one species in a class, then this word was added to the template and the frequencies at which it occurred in all other species of the class were calculated and stored in the database. Classes represented by many species have larger templates of signature words as every species contributes to the template by adding its own species-specific words. It is expected, however, that with adding more species to a class template the template size eventually will reach a saturation limit when adding new species will not increase the template size.

To identify the optimal DS threshold the size of the class templates and the sharing of the signature words between the templates were analyzed at the DS threshold values of 0.1, 0.3 and 0.5 (Fig. 1). The two curves shown in Fig. 1 depict the total number of words of different length in all 22 class templates and the numbers of non-redundant words calculated after removing of all duplications from the combined list of signature words.

As it was expected, the specificity of the words increases with the word length. However, it was noted that the curve of non-redundant word numbers has two extreme maxima at 9 and 14mers with a minimum at 12mers. This implies the existence of two categories of overrepresented oligomers of 9 and 14 (or longer) nucleotides, and the intermediate 12mers mostly occurring as expected. The DS threshold of 0.3 ensures the proportional presentation of word categories in the class templates and the optimal ratio between template size and word specificity.

A list of the 100 most common words present in templates of different classes is given in supplementary Table S2. The most common word that is overrepresented in the genomes of 17 different classes is AAAATTTT (direct and reverse complements are combined and summed). This word is abundant in genomes of all bacterial classes except Acidobacteria, Actinobacteria, Deinococcus/Thermus, Planctomycetes and Thermotogae.

The total occurrence of all signature words per genome was proportional to the length of the genome and was on average 8.8 signature words per 10 bp. (Note that the database is highly redundant. Thus, the 14mer AGGTCTTATGCGGA may be expressed as two 13mers AGGTCTTATGCGG and GGTCTTATGCGGA; three 12mers; four 11mers and so on down to seven 8mers. If each of these words is overrepresented in a genome, the initial sequence of 14 nucleotides contains 28 signature words. The redundancy of the database allows for the selection of the most discriminative signatures.)

The number of expected signature words per 10 bp of the genome sequence varies from 0.7 (*Dehalococcoides* sp. CBDB1, class Chloroflexi) to 23.4 (Candidatus *Carsonella ruddii* PV, class γ -Proteobacteria). In general the signature words were relatively infrequent in α -Proteobacteria (with the exception of the genus *Rhodopseudomonas*),

Bacteroidetes/Chlorobi, Chlamydiae/Verrucomicrobia, Chloroflexi, Crenarchaeota, Cyanobacteria, Euryarchaeota, Nanoarchaeota, Planctomycetes, Thermotogae and the genus *Magnetococcus* (unclassified Proteobacteria). In contrast signature words were abundant in Actinobacteria, β -Proteobacteria, Deinococcus/Thermus, several Firmicutes (genus *Mycoplasma*) and γ -Proteobacteria. The average, maxima and minima of signature words in genomes of different classes are shown in Fig. 2 (see as well Table S1).

Biases to coding regions

The biased distribution of words between coding and noncoding genomic sequences was measured by χ^2 statistics (see ‘Experimental procedures’). In Table S1 words from each genome were classified as coding or noncoding if their distribution was consistently skewed to coding or noncoding genomic regions (p -value 0.001). In the majority of genomes repeated elements are biased exclusively to the coding regions. In our previous study it was demonstrated that the majority of highly overrepresented repeats in coding sequences occur predominantly in the reading frame, e.g. they encode real amino acid words.

Non-coding repeats are usually longer than coding repeats and often are part of longer repetitive nucleotide motifs (Davenport *et al.*, 2009). Non-coding repeats were not detected at the DS threshold of 0.1 in Acidobacteria, Deinococcus/Thermus, ϵ - Proteobacteria, Euryarchaeota, Nanoarchaeota and *Magnetococcus*. However, the genomes *Streptomyces coelicolor* A3(2) (Actinobacteria) and *Erwinia carotovora* subsp. *atroseptica* SCRI1043 (γ -Proteobacteria) showed highly irregular oligonucleotide

patterns as compared to other bacteria with multiple signature words biased to the non-coding regions. Alignment of the signature words showed that in *S. coelicolor* A3(2) genome the motifs CTCGGGTCGGCCACAAGGCACCGT, TTYGGATCAGGCCGGCCCCGGCCG, GGGTCGGCCACAAGGCACCGTCGGCCGCGGCCG, GGCGGGGAGAGCCGCGTACGGCGAGGAGGGGGC, GGGTCGCGCACAAAGGCACCGTCGGCCGCGGCCGGCCTGATCCGAACGACACC, GCTCTCCCCGCCATCGCTCCGGGTAC are repeated multiple times in non-coding regions. *E. carotovora* harbours a more versatile repertoire of non-coding repeats (see the alignment files *S.coelicolor.aln* and *E.carotovora.aln* in supplementary materials).

Using the set χ^2 -threshold an average of 64% of words was not biased to coding or non-coding regions of the genome. The signature words were distributed almost randomly in *Bifidobacterium*, many *Corynebacterium*, *Tropheryma*, *Bartonella*, *Brucella*, *Ehrlichia*, *Rickettsia*, *Chlamydia*, *Chlamydophila*, *Mycoplasma*, *Streptococcus*, *Buchnera*, *Blochmannia*, *Francisella*, *Haemophilus*, *Borrelia* and *Treponema*, many of which are obligate intracellular pathogens or symbionts with simplified metabolism and reduced genome size. Contrary, the signature words are biased to either coding or non-coding regions in *Acidobacteria*, *Frankia*, *S. coelicolor* (but not *S. avermitilis*), *Rhodopseudomonas*, *Aquifex*, *Bacteroides*, *Acidovorax*, *Bordetella*, *Dehalococcoides*, *Myxococcus*, *Bacillus*, *Erwinia*, *Escherichia*, *Pseudomonas*, *Xanthomonas*, *Nanoarchaeum* and *Rhodopirellula* (see Table S1). Consider for example one extreme pattern in *Frankia alni* ACN14a (Actinobacteria) with 13,799,417 copies of 35,382 signature words, 99% of which are biased to the coding regions. Many of these words are

parts of repeated coding motifs (see the alignment file F.alni.aln in the supplementary materials) which often encode the arginine-rich low complexity regions in the protein sequences of this organism.

An investigation was undertaken to determine whether the oligonucleotides found predominantly in coding or non-coding regions of one genome show the same preferences in other genomes. The distribution of the most common word AAAATTTT (see above) was studied. This word was identified with a DS value over 0.5 in 169 genomes. In 150 genomes it was found predominantly in coding sequences, in 9 genomes this word appeared abundantly in non-coding sequences, and in 10 genomes the distribution was statistically unbiased (see supplementary Table S3). Next, the distribution of signature words of *E. carotovora* subsp. *atroseptica* SCRI1043 (all signature words from this genome are associated with non-coding sequences, see above) was analyzed (supplementary Table S4). Most of these words were unique to this strain or shared among a few other representatives of the class γ -Proteobacteria, mostly with *Photobacterium luminescens* subsp. *laumondii* TTO1. However, in other genomes these words were found in both coding and non-coding sequences. The word ACCAACTGAGCTA is overrepresented at the DS level of 0.5 in 40 species of γ -Proteobacteria, thus it may serve as a marker for this class. To summarize, the words which are found to be biased in one genome generally cannot be expected to show similar biases in other bacteria.

Taxonomic specificity of the signature words

Scoring algorithm was used to determine the ability of the various overrepresented words to identify which species, genus or class a genomic fragment is from. The performance of

the words depends on the ultimate purpose of the identification and on the identification approach that is used. In general, non-coding words are longer and more specific making them better suited for monitoring species of interest within a habitat. Profiles of oligonucleotides which are abundant in coding sequences may be a better choice for whole species identification in a metagenomic sample.

To experimentally test the discriminative power of selected words several pseudo-metagenomic datasets were generated by the program MetaSim from completely sequenced bacterial chromosomes (Richter *et al.*, 2008). Several algorithms have been developed to score the signature words.

In order to identify whether an oligonucleotide is species specific a score was assigned based on its ability to distinguish one species from another within the dataset. A ratio based approach was selected to calculate a Species Specificity Score (SSS) for a given word for the species i in the set of N species:

$$SSS = \frac{DS_i \times \left\{ \sum_{j \neq i}^N \frac{DS_i}{DS_j + 1} \right\}}{N} \quad (2)$$

To avoid redundancy of selected words, highly similar oligonucleotides were filtered out. A progressive heuristic approach was used to check subsequences within each oligonucleotide against all other oligonucleotides. Oligonucleotides were identified as highly similar if subsequences matched and similar words with lower scores were removed from the list of signature words.

Consider the following examples. *Pseudomonas aeruginosa* could be easily differentiated from its relatives by the 10 top-scoring species-specific 9mers (Table 1). However, *P. putida* cannot be differentiated by 9mer as the words biased to the coding regions of

the organism are not specific and the longer non-coding 13mers should be used to distinguish *P. putida* from the closely related species (Table 2). Similarly, *Mycobacterium tuberculosis* may not be distinguished from the closely related species *M. bovis* and *M. marinum*, which harbor very similar numbers of signature words to *M. tuberculosis* (Table 3). In the *M. bovis* genome there is an average of 6 signature words per 10 bp, while the frequency of signature words in the *M. tuberculosis* genome is 2 per 10 bp (Table S1). This explains why the total count of species specific oligonucleotides selected for *M. tuberculosis* may be the same or higher for genomic fragments generated from *M. bovis*. In this scenario strain or species-specific non-coding marker words may provide a solution. For example in Table 2 a 16mer GATCGCAAGCGCGGCG was used to separate *M. tuberculosis* from *M. bovis*. Although this work was initially focused on the analysis of 8 to 14mers, the availability in the database of coordinates for all word instances per genome makes it possible to enlarge, concatenate and align the overlapping words to get the statistics of the distribution of longer sequences.

In order to identify oligonucleotides which occur abundantly within a single phylogenetic lineage (genus, family or class) two criteria needed to be satisfied. Firstly, the most generally overrepresented oligonucleotides for this lineage were identified. This was achieved by calculating a Lineage Commonality Score (LCS, see equation 3).

$$LCS = \frac{DS'}{\sigma^2 + 1} \quad (3)$$

where DS' is the average DS value for this word for the whole lineage and σ^2 is the variance of the DS values.

The LCS may also be applied to order or filter the list of prospective signature words. The second criterion was to determine whether the most commonly overrepresented oligonucleotides are specific for this lineage by calculating a Lineage Specificity Scores (LSS, see equation 4). The LSS equation resembles the function used to calculate the species specific scores (equation 2) but uses the average DS values for all classes or all genera of one class to determine the most descriptive oligomers.

$$LSS = \frac{DS'_i \times \left(\sum_{j \neq i}^N \frac{DS'_i}{DS'_j + 1} \right)}{N} \quad (4)$$

In equation 4, i denotes the class or genus and N is the set of all classes or the set of all genera for the specific class. In Table 4 the signature words selected for the genus *Salmonella* were tested against datasets of chromosomal fragments generated from the genomes of species from several other genera. The genome *S. enterica* subsp. *enterica* serovar Typhi str. CT18 scored higher with *Salmonella*-specific oligomers than the representatives of other species. However, *Klebsiella pneumoniae* achieved almost the same score as *Salmonella* due to the higher frequency of repeated DNA motifs in this genome. These species may be distinguished through the use of more specific non-coding oligonucleotides.

Discussion

In this paper a large scale analysis of the distribution of overrepresented 8 to 14 bp oligomers is presented. We attempted to address whether the distribution of overrepresented words mentioned in our previous work by Davenport *et al.*, 2009 is

common for all microorganisms and if these oligomers can be utilized for species identification of DNA reads in environmental samples.

It was found that the presence of highly repetitive DNA motifs in coding sequences of bacterial genomes encoding di- or tripeptides is a common phenomenon for all bacteria including Archaea. It is possible that these repeats contribute to filling low complexity regions in bacterial proteins. It makes biological sense for bacteria to utilize biochemically neutral motifs rather than random amino acids in large regions between typically conserved structural cores and active sites (Valencia and Pazos, 2003).

However, it is unknown how the bacterial cell may control proliferation of these motifs and if this is controlled by specific molecular mechanisms.

Abundance of repetitive DNA words in non-coding sequences is not widespread among bacteria. Some strains with overrepresented non-coding motifs may be found in many bacterial classes; however, their nearest relatives may be devoid of such repeats (Table S1). In several cases the presence of multiple repeats in non-coding genomic regions is strain-specific, for example in the *Salmonella enterica* genomes. Analysis of the distribution of non-coding repeats is hampered by the fact that identification of this type of signature oligonucleotides relies on the quality of genome annotation. For example, two genomes of *Ehrlichia ruminantium* str. Welgevonden have been sequenced one at the University of Pretoria, RSA (NC_005295) and another one by CIRAD, France (NC_006832). Both of the genomes are of a similar length and contain the same number of signature words (see Table S1) but in NC_005295 these words were associated predominantly with non-coding sequences while in NC_006832 most of them were biased to the predicted ORFs. This discrepancy did definitely not result from biological

differences but rather from the different annotation approaches employed. The above mentioned discrepancy implies that the non-coding words in one genome may be found in coding sequences of another genome because of difference in annotation. However, the differences in distribution of coding and non-coding oligonucleotides discussed in this paper cannot be solely attributed to annotation problems and these two types of repeats may be distinguished by several characteristics. For example, the repeated motifs in non-coding regions usually are longer and more specific for the genome than those in coding regions.

It may be expected that the most commonly distributed words (Table S2) are in the stretches of DNA, which exert less structural constraints and which do not carry any binding sites for regulatory proteins. The most widely distributed motifs are poly-A alternating with poly-T and stretches of (GC)_n. In contrast, (AT)_n, poly-G and poly-C are rarely part of the overrepresented words. This finding do not support the hypothesis of structurally unbiased motifs as the base stacking energy reaches the extreme value in (GC)_n and the stretches of poly-A with poly-T exhibit an extreme propeller twist angle (Baldi and Baisne, 2000). Reservation of TATA-boxes in bacterial genomes for promoter regions may explain the rarity of (AT)_n motifs. An analysis of rare words showed that the majority contained the CTAG tetramer – a stop codon in both direct and reverse complement DNA strands (data not shown).

The biological function of globally overrepresented oligomers remains unclear. In many cases if one organism harbors abundant repeat motifs, its closest relative may be repeat poor (note the example of *M. bovis* and *M. tuberculosis* discussed above). This implies that the abundance of repeats in a genome may result from uncontrolled proliferation of

selfish DNA elements that have no value for the host organism. On the other hand, free living bacteria with versatile metabolic processes and a complex lifestyle are more likely to have globally distributed repeated elements than parasitic and symbiotic microorganisms with a restricted metabolism and a sequestered habitat (see Fig. 2).

The focus of this paper was to study the potential use of signature words for the identification of species in environmental samples and to develop scoring algorithms for tailor-made sets of marker words for the phylogenetic identification at class, genus and species level. Overrepresented 8 to 14bp oligonucleotides are present throughout the genome and can be effectively employed in the identification of bacterial species in metagenomic sequences without the need for sequence assembly. In order to identify oligonucleotide markers raw data on the distribution of overrepresented words were analyzed and imported into a database. The creation of a structured database from raw data offers the opportunity to search and compare data based on selected criteria. The analysis of the distribution of selected words in the class templates demonstrated that the most discriminative words may be found among 9mers and 13-14mers (Fig. 1). Several algorithms for the selection of species and lineage specific signature words have been proposed and tested. There is potential for discovery of various uses for these oligonucleotides including the enhancement of existing identification pipelines which may lead to improved detection methods within metagenomics. Particularly, the identification of unknown DNA sequences may be preceded by clustering of DNA reads using shorter oligonucleotide patterns (McHardy *et al.*, 2007; Nasser *et al.*, 2008). The identification procedure should comprise several iterative processes counting signature words from class, genus and species levels of specificity. This should be followed by

rectifying false-positive results similar to those that were discussed above (where *M. bovis* and *K. pneumoniae* outscored the set of markers designed for other organisms) by using strain and species-specific words. A graphical user interface and the program for identification of unknown DNA reads in metagenomic datasets are now under development and will be presented in further publications.

Experimental procedures

The in-house program OligoCounter (available at <http://webhost1.mh-hannover.de/davenport/oligocounter/index.html>) was utilized in the discovery of overrepresented oligonucleotides from unannotated FASTA files of 438 bacterial genomes (available for download from the NCBI database at <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>) as it was described previously (Davenport *et al.*, 2009). Raw data from OligoCounter was used to populate the developed MySQL database.

To study the biased occurrence of overrepresented oligonucleotides between coding and non-coding sequences their distribution was checked against a null-hypothesis of the random distribution by the χ^2 criterion (Fleiss *et al.*, 2003):

$$\chi^2 = \frac{2 \left(|N_c E_{nc} - N_{nc} E_c| - N \right)^2}{N (N_c + E_c) (N_{nc} + E_{nc})} \quad (5)$$

where N is the number of all instances of a word in a genome; N_c – number of coding words; N_{nc} – number of non-coding words; E_c – number of coding words expected from an assumption of the random distribution of this word; E_{nc} – number of expected non-coding words (it should be clear that $N = N_c + N_{nc} = E_c + E_{nc}$). Occurrence of the words

in coding and non-coding sequences was determined by referring to the genome annotation data in GenBank files from the NCBI database (Benson *et al.*, 2007).

The expected numbers of oligonucleotide words in coding and non-coding sequences were calculated based on the assumption that these numbers should be proportional to the ratio of lengths of coding and non-coding sequences in this genome. As this ratio was strongly affected by the quality of annotation rather than by natural biological reasons (Ussery and Hallin, 2004), instead of calculating the coding/non-coding sequence ratio for every genome a modal ratio of 88% of coding sequences per genome was used. This ratio was calculated by the annotation information of 684 completely sequences bacterial genomes available now at the NCBI (see Fig. S1 in supplementary materials).

Thus, E_c and E_{nc} values in above mentioned calculations were: $E_c = 0.88N$ and $E_{nc} = 0.12N$. The distribution of a word was defined as biased if χ^2 was 10.83, which corresponds to a p -value of 0.001 (Fleiss *et al.*, 2003).

To concatenate predicted overrepresented words into longer repetitive motifs an in-house Python program was used that has previously been tested in identification of multiple repeats in bacterial giant genes (Reva and Tümmler, 2008). Alignment of repeats was done using the ClustalW algorithm in the program JalView (Waterhouse *et al.*, 2009).

To generate pseudo-metagenomic datasets the MetaSim program was used (Richter *et al.*, 2008). The DNA clones size was set to follow a uniform distribution between 800bp and 1300bp. For every genome a set of 100 artificial DNA reads with an average length of 1Mb was generated.

Acknowledgements

This research was supported by a grant of the National Bioinformatics Network of South Africa. C.F.D. was sponsored by the Deutsche Forschungsgemeinschaft as a member of the International Research Training Group “Pseudomonas Pathogenicity and Biotechnology” (GRK 653/3).

References

- Baldi, P. and Baisnée, P.-F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all length. *Bioinformatics* **16**: 865-889.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2007) Genbank. *Nucleic Acids Res* **35**: D21–D25.
- Chang, C.-H., Chang, Y.-C., Underwood, A., Chiou, C.-S. and Kao, C.-Y. (2007) VNTRDB: a bacterial variable number tandem repeat locus database. *Nucleic Acids Res* **35**: D416–D421.
- Davenport, C. F., Wiehlmann, L., Reva, O. N. and Tümmler, B. (2009) Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ Microbiol* [Epub ahead of print].
- DeLong, E. F. (2005) Microbial community genomics in the ocean. *Nat Rev Microbiol* **3**: 459-469.
- Dimri, G. P., Rudd, K. E., Morgan, M. K., Bayat, H. and Ames, G. F. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and

phylogenetic distribution among *Escherichia coli* strains and other enteric bacteria.
J Bacteriol **174**: 4583–4593.

Fleiss, J. L., Levin, B., Cho Paik, M. (2003) Statistical methods for rates and proportions.
3rd Edition. New York, USA: Wiley.

Hulton, C. S., Higgins, C. F. and Sharp, P. M. (1991) ERIC sequences: a novel family of
repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium*
and other enterobacteria. *Mol Microbiol* **5**: 825–834.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity.
Curr Opin Microbiol **1**: 598–610.

Mancuso, M., Avendaño-Herrera, R., Zacccone, R., Toranzo, A. E. and Magariños, B.
(2007) Evaluation of different DNA-based fingerprinting methods for typing
Photobacterium damsela ssp. *piscicida*. *Biol Res* **40**: 85–92.

McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. and Rigoutsos, I. (2007)
Accurate phylogenetic classification of variable-length DNA fragments. *Nat*
Methods **4**: 63-72.

Mojica, F. J., Díez-Villaseñor, C. and Soria, E. (2005) Intervening sequences of regularly
spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**:
174-182.

Mojica, F. J., Díez-Villaseñor, C., Soria, E. and Juez, G. (2000) Biological significance of
a family of regularly spaced repeats in the genomes of archaea, bacteria and
mitochondria. *Mol Microbiol* **36**: 244–246.

Mongodin, E. F., Emerson, J. B. and Nelson, K. E. (2005) Microbial metagenomics.
Genome Biol **6**: 347.

- Nasser, S., Breland, A., Harris, F. C. Jr. and Nicolescu, M. (2009) A fuzzy classifier to taxonomically group DNA fragments within a metagenome. *IEEE Xplore* 1-6.
- Reva, O. and Tümmler, B. (2008) Think big – giant genes in bacteria. *Environ Microbiol* **10**: 768-777.
- Richter, D. C., Ott, F., Auch, A. F., Schmid, R. and Huson, D. H. (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE* **3**: e3373.
- Rodriguez-Valera, F. (2002) Approaches to prokaryotic biodiversity: a population genetics. *Environ Microbiol* **4**: 628-633.
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**: 163.
- Ussery, D. W. and Hallin, P. F. (2004) Genome Update: annotation quality in sequenced microbial genomes. *Microbiology* **150**: 2015-2017.
- Valencia, A. and Pazos, F. (2003) Prediction of protein-protein interactions from evolutionary information. In: Structural Bioinformatics, Bourne, P.E and Weissig, H. (eds.), Wiley-Liss, Hoboken, NJ, USA: 501-507.
- van Belkum, A. (2007) Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *FEMS Immunol Med Microbiol* **49**: 22-27.
- Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. and Barton, G. J. (2009) Jalview version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* [Epub ahead of print].

Tables

Table 1. Counts of the top 10 signature 9mers selected for *P. aeruginosa* in artificial DNA reads (1 Mb on average) generated from the *Pseudomonas* genomes.

Words	SSS	NC_002516	NC_004129	NC_002947	NC_008027	NC_004578
GCCGCCGGC	0.34	17	7	10	1	0
GCCGGCGCC	0.29	19	9	7	6	0
GCCGAGCAG	0.25	22	7	4	5	5
GCCAGGCGC	0.22	16	7	6	8	6
CCAGGCCGA	0.22	19	11	7	9	2
CTCGGCCTG	0.22	11	10	2	9	3
TGCTCGGCG	0.22	17	6	1	7	2
GCGCGGCGA	0.21	8	3	4	3	5
CCTGCTGGC	0.18	10	7	6	11	2
CTGCTGCTG	0.18	9	15	8	6	3
Accumulated counts:		148	82	55	65	28

NC_002516 – *P. aeruginosa* PAO1; NC_004129 – *P. fluorescens Pf-5*; NC_002947 – *P. putida* KT2440; NC_008027 – *P. entomophila* L48; NC_004578 – *P. syringae* pv. tomato DC3000. The words are sorted by SSS values. For attribution of DNA reads the accumulated counts are calculated.

Table 2. Counts of the top 10 signature 9mers and the top 10 signature 13 to 14mers selected for *P. putida* in artificial DNA reads (1 Mb on average) generated from the *Pseudomonas* genomes.

Words	SSS	NC_002947	NC_008027	NC_004129	NC_002516	NC_004578
CTGCTGGCC	0.16	10	15	9	9	2
CAGCGCCTG	0.15	9	9	5	15	7
TGCTGCTGG	0.15	11	10	11	10	2
GGCGCTGGC	0.14	6	10	7	10	5
GGCCAGGCC	0.13	10	5	6	7	3
CCAGCAGGC	0.13	7	7	8	13	3
CTGGCCAGC	0.13	9	12	6	7	7
GCCTGCTGC	0.12	5	13	8	12	8
ACCCGCGAA	0.11	3	2	0	0	2
GCAGGCGCT	0.11	12	5	6	10	3
Accumulated counts:		82	88	66	93	42
GGCCTCTTCGCGG	0.06	2	0	0	0	0
CCGCTCCCACAGG	0.04	1	3	0	0	0
CGGCCTCTTCGCG	0.03	1	0	0	0	0
TTCGCGGGTAAAC	0.03	1	1	0	0	0
TCGCGGGTAAACC	0.03	1	1	0	0	0
GTTCACCCGCGAA	0.03	0	0	0	0	0
GGGCTTGCCCGCG	0.03	1	0	0	0	0
AGCGGGTTTACCC	0.02	2	1	0	0	0
GTAAACCCGCTCC	0.02	1	1	0	0	0
CTGTGGGAGCGGGC	0.02	3	0	0	0	0
Accumulated counts:		13	7	0	0	0

NC_002947 – *P. putida* KT2440; NC_002516 – *P. aeruginosa* PAO1; NC_004129 – *P. fluorescens* Pf-5; NC_008027 – *P. entomophila* L48; NC_004578 – *P. syringae* pv. tomato DC3000. The words are sorted by SSS values. For attribution of DNA reads the accumulated counts are calculated.

Table 3. Counts of the top 9 signature 9mers, one 14mer and one 16mer marker selected for *M. tuberculosis* in artificial DNA reads (1 Mb on average) generated from *Mycobacteria*.

Words	SSS	NC_00096 2	NC_00876 9	NC_01061 2	NC_00859 6	NC_00294 4	NC_00872 6	NC_01039 7	NC_00267 7
GGCGGCAACG	0.10	2	8	6	1	2	3	0	0
TGGCCGCGGC	0.06	2	3	1	1	2	0	2	1
CGGTGGCGCC	0.06	5	3	1	2	0	4	2	1
TTGGCCGCCG	0.05	3	3	5	0	2	0	0	3
TGCTGGCCGG	0.05	2	2	4	1	3	3	0	1
CGTCACCGCC	0.05	6	3	4	8	0	4	1	0
GGCGATCACC	0.04	2	0	3	4	8	1	3	1
TCGGCCAGCA	0.04	3	1	0	3	1	2	0	2
ACCGCCGGCG	0.04	2	2	1	2	1	0	0	0
GGCGGGGCCGGC GG	0.04	4	1	2	0	1	0	0	0
Accumulated counts:		31	26	27	22	20	17	8	9
<i>M. tuberculosis</i> unique marker word GATCGCAAGCGCGGCG		2	0	0	0	0	0	0	0

NC_000962 – *M. tuberculosis* H37Rv; NC_008769 – *M. bovis* BCG str. Pasteur 1173P2; NC_010612 – *M. marinum* M;
NC_008596 – *M. smegmatis* str. MC2 155; NC_002944 – *M. avium* subsp. *paratuberculosis* K-10; NC_008726 – *M. vanbaalenii*
PYP-1; NC_010397 – *M. abscessus*; NC_002677 – *M. leprae* TN. The words are sorted by SSS values. For attribution of DNA reads
the accumulated counts are calculated.

Table 4. Counts of the top 10 signature 8mers selected for *Salmonella* in artificial DNA reads (1 Mb on average) generated from the genomes of different bacterial classes.

Words	LSS	NC_00319	NC_01128	NC_00091	NC_00251	NC_00314	NC_00250	NC_00096	NC_00092
		8	3	3	6	3	5	4	1
GCCGGAAA	45.1	9	2	6	5	5	4	3	0
CGGCGATA	35.7	13	6	6	3	6	1	1	0
TCGCGCTG	33.3	7	8	2	9	3	1	0	2
CCAGACGC	31.4	6	5	5	4	0	1	1	0
GCCGGAAG	30.4	9	4	6	6	1	1	4	0
ATCCGGCG	29.2	10	7	4	6	2	1	1	1
GCGTACCG	28.9	6	7	4	5	2	3	0	0
GCCAGACG	28.3	8	3	5	2	0	1	3	0
TTAACGCC	27.1	6	5	8	0	4	1	2	6
TCGCCAGC	27.0	6	6	2	8	1	0	1	0
G									
Accumulated counts:		80	53	48	48	24	14	16	9

NC_003198 – *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18; – *Klebsiella pneumoniae* 342; NC_000913 – *Escherichia coli* K12; NC_002516 – *Pseudomonas aeruginosa* PAO1; NC_003143 – *Yersinia pestis* CO92; NC_002505 – *Vibrio cholerae* O1 chromosome I; NC_000964 – *Bacillus subtilis* subsp. *subtilis* str. 168; NC_000921 – *Helicobacter pylori* J99. The words are sorted by LSS values. For attribution of DNA reads the accumulated counts are calculated.

Figure legends

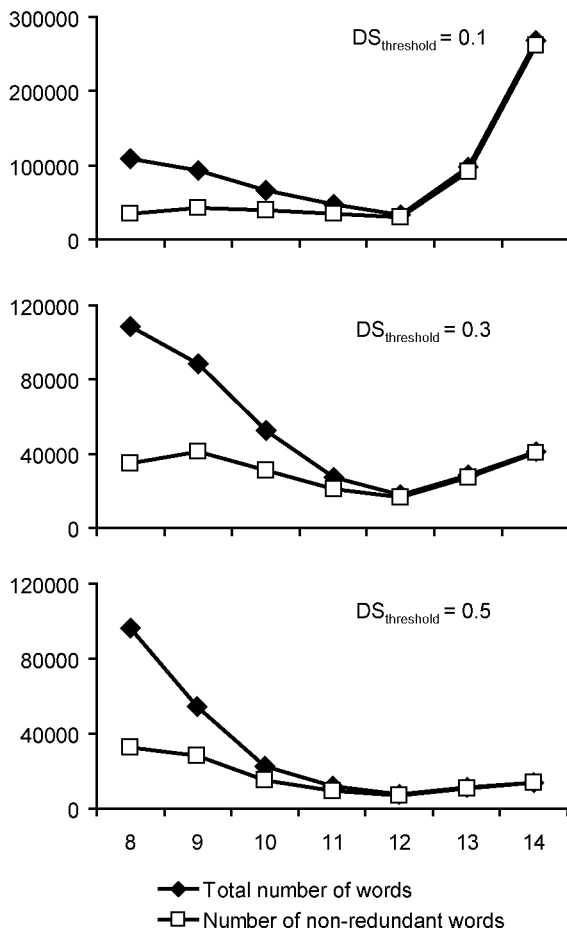


Fig. 1. The distribution of words of different lengths selected at the DS threshold values of 0.1, 0.3 and 0.5. The line marked with closed diamonds shows the total size of the signature word templates generated for all classes, while that with open boxes shows the size of the combined template after deletion of all duplicated words.

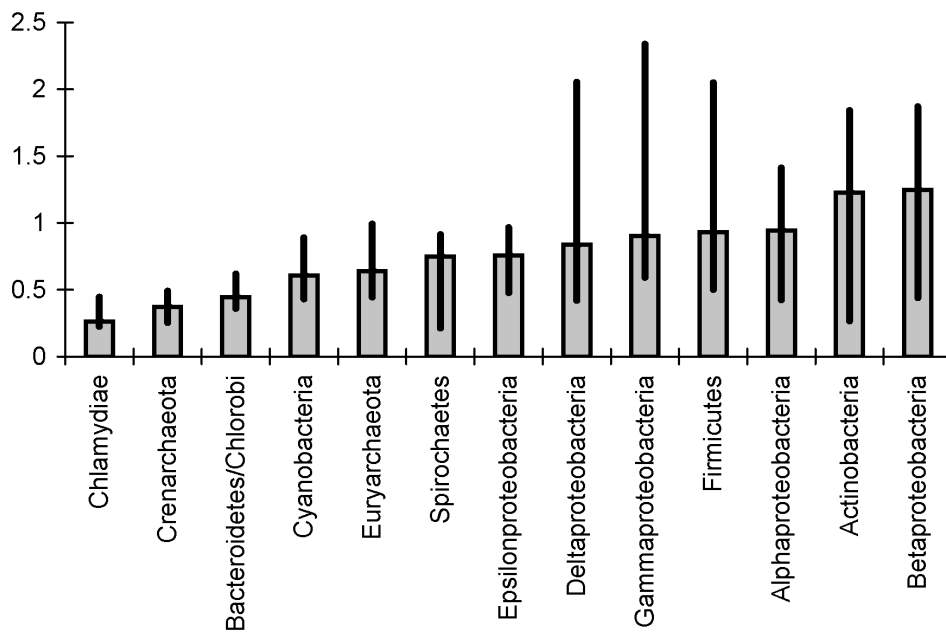


Fig. 2. The average, maximal and minimal numbers of signature words in genomes of different classes. The columns indicate the average values and the lines show the range of variation of the density of signature words per class.

Supporting information

Additional supporting information may be found in the online version of this article:

Table S1. Signature words identified in bacterial genomes with the DS-threshold of 0.1.

Table S2. Word sharing in the templates of bacterial classes. The top 100 words were selected with the DS threshold of 0.1. The words were sorted by the number of templates where these words were present.

Table S3. Distribution of the most common word AAAATTTT in bacterial genomes.

Only the genomes are listed where this word is distributed above the DS-threshold of 0.5.

Table S4. The oligomers which are overrepresented in *Erwinia carotovora* subsp.

atroseptica SCRI1043

E.carotovora.aln. An alignment of signature words of *Erwinia carotovora* subsp. *atroseptica* SCRI1043.

F.alni.aln. An alignment of signature words of *Frankia alni* ACN14a.

S.coelicolor.aln. An alignment of signature words of *Streptomyces coelicolor* A3(2).

Fig. S1. Distribution of the coding sequences percentage in 684 completely sequenced bacterial chromosomes.

4. Genome Diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains

4.1 Background

This laboratory has maintained a large collection of bacterial strains for many years. These strains were isolated in the past, some as long ago as the 1940s, and have been passed on to different laboratories across the world. Various events such as subcloning, contamination or maintenance at inappropriate temperatures have putatively driven microevolution in these sublines. Microevolution is already a known problem in microbiology laboratories (Spira et al. 2008). Most probably, the extent of genomic modifications has likely been underestimated to date given the lack of tools to examine these with. One example is hypermutator strains. These strains, which have suffered mutations in their genomic mutation repair genes such as *mutT* or *mutS* (Woodford and Ellington 2007, Winstanley et al. 2009), then accumulate mutations in the rest of their genome very quickly and typically display a highly abnormal phenotype. Hypermutators are in most cases unsuitable for future research. Should sequencing costs continue to decrease isolates in microbiological strain databanks may in future be routinely sequenced to better quality control.

4.1.1 Genome assembly with short reads

Genome assembly, the art of clustering and positioning reads into large contiguous stretches of sequence termed contigs, has always been a core element of bioinformatics. Two variant methodologies are recognised: *de novo* or reference assisted assemblies. Both were required in this *P. aeruginosa* resequencing project. *De novo* assembly assumes no prior information other than that in the reads to build contigs. In most cases this is done via an overlap-layout-consensus algorithm (Pop et al. 2004). First reads with some similarity are aligned to one another to generate overlapping reads, then the layout step positions reads with respect to one another resulting in a multiple alignment of all reads, and finally a consensus sequence based on the most common base at each position from the numerous reads is produced. Popular algorithms which use this system are Phrap (Ewing and Green 1998), Minimus (Sommer et al. 2007) and Edena (Hernandez et al. 2008). Reference assembly is dependent upon the availability of a closely related reference sequence. Reads are then mapped onto the reference and a consensus is generated. This approach is relatively rapid and computationally undemanding. However regions not in the reference sequence will be simply ignored. Repeats, which are a serious, long term and unavoidable problem in sequence assembly, complicate matters in both scenarios. Reads mapping to multiple positions in a reference

genome are typically mapped randomly to one of the two (Li and Durbin 2009). In contrast, contigs created by *de novo* assemblers cannot be continued beyond a repeat (Zerbino and Birney 2008, Pop and Salzberg 2008). Repeats are an even more serious problem in short read assembly (Zerbino and Birney 2008). This is because assembly programs experience the aforementioned problems when repeat length exceeds read length. When using short 36 bp reads, as in this manuscript, many more common short repeat regions limit assemblies which would otherwise be spanned by a much longer read from a traditional Sanger sequencing project.

Modern short read resequencing projects commonly have an initial reference assembly phase followed by a *de novo* assembly phase. In this second phase regions of DNA which are putatively unique to the organism being sequenced are identified by assembly of reads not mapped to the reference sequence. The closer the genome under study to the reference genome, the fewer reads will be from novel DNA. Care must be taken to discern between missassemblies created by the *de novo* alignment algorithm and high confidence contigs which are truly novel genomic regions. Typically a conservative coverage cutoff of 10 reads is also used here to filter out low confidence data.

4.2 About the manuscript

This manuscript follows the investigation of two *Pseudomonas aeruginosa* strain PAO1 sublines, termed MPAO1 and PAO-DSM. These sublines have been present in various laboratories for a number of years after isolation in Australia in the 1940s and differences have long been apparent. Unpublished wet laboratory data had already been collected which illustrate the phenotypic anomalies of these sublines. Thorough analysis of these data was only feasible with a resequencing approach to discover their genetic basis. As these strains are only likely to be subtly different to the already sequenced PAO1 (Stover et al. 2000), a sequencing technology able to reliably detect SNPs was required. Over ten million short 36 bp reads from an Illumina device were thus used. The main differences from PAO1 of the two sublines include a missing large inversion, deletions of between 3 and 1006 bp in length, and 39 single nucleotide polymorphisms, 17 of which affect protein sequences. Examples include substitutions in the FtsZ cell division protein and NapA nitrate reductase, which is involved in anaerobiosis. A prophage was also implicated by *de novo* assembly of reads into five contigs, some of which span the phage cargo region containing phosphotransferases and kinases. On the phenotypic level, in a murine airway infection model PAO1 sublines responded differently to nutrient limitation and heat stress, also showing modified

virulence. A PAO-DSM subpopulation also displayed an arginine to leucine substitution in 9/79 reads in the LasR locus, leading to defective quorum sensing and pleiotropic phenotypic effects in a subpopulation. This observation exemplifies the sensitivity of the Illumina Genome Analyzer system. Lastly, results from competition experiments showed a clear advantage for the original PAO1 and subline MPAO1, which immediately killed PAO-DSM cells when they were grouped in a mixed population. In summary, the finding that various *P. aeruginosa* strains can evolve despite precautions in the laboratory environment has ramifications for the reproducibility of research on apparently identical model organisms.

This manuscript is in revision at the Journal of Bacteriology at the time of writing. My contribution was restricted to the establishment of a short read assembly pipeline, bioinformatic assistance and annotation of the novel phage DNA.

4.3 References

- Ewing, B. & Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**(3) 186-194.
- Hernandez, D., François, P., Farinelli, L., Osterås, M. & Schrenzel, J. (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* **18**(5) 802-809.
- Li, H. & Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**(14) 1754-60.
- Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. (2004) Comparative genome assembly. *Brief Bioinform* **5**(3) 237-248.
- Pop, M. & Salzberg, S. L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet* **24**(3) 142-149.
- Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. (2007) Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**, 64.
- Spira, B., Hu, X. & Ferenci, T. (2008) Strain variation in ppGpp concentration and RpoS levels in laboratory strains of *Escherichia coli* K-12. *Microbiology* **154**(Pt 9) 2887-2895.
- Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrenner, P., Hickey, M. J., Brinkman, F. S., et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**(6799) 959-964.

Winstanley, C., Langille, M. G. I., 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**(1) 12-23.

Woodford, N. & Ellington, M. J. (2007) The emergence of antibiotic resistance by mutation. *Clin Microbiol Infect* **13**(1) 5-18.

Zerbino, D. R. & Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**(5) 821-829.

Genome Diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains

Jens Klockgether^{1*}, Antje Munder¹, Jens Neugebauer¹, Colin F. Davenport¹, Frauke Stanke¹,
Karen D. Larbig¹, Stephan Heeb², Ulrike Schöck³, Thomas M. Pohl³, Lutz Wiehlmann¹,
Burkhard Tümmler¹

¹Klinische Forschergruppe, OE 6710, Medizinische Hochschule Hannover, D-30625
Hannover, Germany

²Centre for Biomolecular Sciences, University of Nottingham, Nottingham NG7 RD, United
Kingdom

³GATC Biotech AG, D-78467 Konstanz, Germany

Running title: *P. aeruginosa* PAO1 genome diversity

Keywords: Genome Analyzer, genome sequencing-by-synthesis, microevolution,
Pseudomonas aeruginosa, pulsed field gel electrophoresis, reference strain

*Corresponding author:

Jens Klockgether, Klinische Forschergruppe, OE 6710, Medizinische Hochschule Hannover,
Carl-Neuberg-Str. 1, D-30625 Hannover, Germany, phone: +49-511-5326721, FAX:
+49-511-5326723, email: klockgether.jens@mh-hannover.de

Abstract

Pseudomonas aeruginosa PAO1 is the most commonly used strain for research on this ubiquitous and metabolically versatile opportunistic pathogen. Strain PAO1, isolated in the 1940s in Australia, has been distributed worldwide to laboratories and strain collections. Meanwhile discordant phenotypes of PAO1 sublines have emerged. Taking the existing PAO1-UW genome sequence as a blueprint, the genome sequences of reference strains MPAO1 and PAO-DSM were resolved by physical mapping and deep short read sequencing-by-synthesis. MPAO1 has been the source of near-saturation libraries of transposon insertion mutants and PAO-DSM is identical in its SpeI-DpnI restriction map with the original isolate. The major genomic differences between MPAO1 and PAO-DSM in comparison to PAO1-UW are the lack of a large inversion, a duplication of a mobile 12 kb prophage region carrying a distinct integrase and protein phosphotransferases or kinases, deletions of 3 to 1006 bp in size and at least 39 single nucleotide substitutions, 17 of which affect protein sequences. The PAO1 sublines differed in their ability to cope with heat stress or nutrient limitation and their virulence in an acute murine airway infection model. Most unexpectedly the MPAO1 and PAO-UW sublines killed PAO-DSM immediately when strains were pooled together into one bacterial population. In conclusion, *P. aeruginosa* PAO1 shows an on-going microevolution of genotype and phenotype that jeopardizes the reproducibility of research. High-throughput genome re-sequencing will resolve more cases and could become a proper quality control for strain collections.

Introduction

The metabolically versatile *Pseudomonas aeruginosa* is an opportunistic pathogen of plants, animals and humans and is ubiquitously distributed in soil and aquatic habitats. The common reference strain is *P. aeruginosa* PAO, later named PAO1. Originally isolated from a burn wound in the 1940s (9), the PAO1 strain has become the reference strain for *Pseudomonas*

genetics and functional analyses of the physiology and metabolism of this gamma-proteobacterium. A genetic map of the PAO1 chromosome was generated by exploiting the mechanisms of gene exchange in bacteria, i.e. transduction, transformation and conjugation (10). With the advent of pulsed field gel electrophoresis, a physical map of the PAO1 genome was constructed (27) and later merged with the genetic map information (11). By 2000 the PAO1 strain had been completely sequenced (31). Thereafter the genome annotation has been continually updated and the database content and functionality has been expanded to facilitate accelerated discovery of *P. aeruginosa* drug targets and vaccine candidates (32). Two near-saturation libraries of transposon insertion mutants have been constructed in *P. aeruginosa* PAO1 as a global resource for the scientific community (13, 19).

Comparison of the genome sequence with the physical map revealed a large 2.2 Mb inversion between the sequenced strain and the original strain that had been previously used for genetic and physical mapping (31). Thus we became suspicious that the PAO1 sublines maintained worldwide in numerous laboratories and strain collections may have meanwhile diversified their genomic sequence by mutational events. Indeed, a PAO1 subline from a German strain collection (PAO1-D) and another, independent PAO1 subline from a Japanese strain collection (PAO1-J) that had been stored by research groups in Germany and Japan, respectively, were found to be quorum-sensing negative mutants that carried point mutations in the regulatory gene *lasR* (6). In addition, spontaneous secretion-defective *vfr* mutants from a PAO1 population were observed after several cycles of static growth (2). Similarly, we noted a difference in virulence in a mouse infection model (see below) between the M-PAO1 and PAO-DSM sublines that had been utilized for the construction of the transposon library (13) and the physical map (27), respectively. PAO-DSM was indistinguishable in its SpeI-DpnI-SwaI-PacI physical map from the PAO1 subline that had been stored in Dr. Holloway's laboratory (11). Hence, we decided to compare the genomic sequence of the initially sequenced PAO1 subline PAO1-UW (31) with that of MPAO1 and of PAO-DSM. Combined

physical mapping and DNA sequencing-by-synthesis revealed numerous SNPs and indels in the chromosomes that were associated with differences in global fitness and virulence of the sublines.

Materials and Methods

Bacterial strains.

PAO-DSM was obtained in 1986 and again in 2004 from the German Collection of Microorganisms and Cell Cultures (DSMZ), Brunswick (catalogue no. DSM 1707). MPAO1 was received in 2003 from the distributor of the PAO1 mutant library of the University of Washington, Seattle (13). PAO1-UW was supplied in 1999 by A.L. Erwin, Pathogenesis Corporation, Seattle, and again in 2005 by T. Rasmussen, DTU, Lyngby. For long term storage glycerol stocks were generated from freshly grown cultures and kept at -80°C .

Media and culturing.

P. aeruginosa PAO1 sublines were cultivated in either LB broth or in M9 medium supplemented with 30 mM succinate. Cells were grown at 37°C if used as a source for DNA preparation. To investigate the killing of PAO-DSM by the other PAO1 sublines, cells of a 2 mL exponential culture of PAO-UW or MPAO1 were pelleted. The supernatant was filtered through a sterile $0.45\ \mu$ filter and 250 μL thereof were dropped onto top agar freshly inoculated with *P. aeruginosa* PAO1-DSM.

Preparation of genomic DNA.

Genomic DNA was prepared from bacterial cells following standard procedures (1). To improve DNA purity additional phenol-chloroform-extraction steps were included.

Physical mapping.

Regions suspicious for insertions or deletions in PAO-DSM and MPAO1 were analysed by partial or complete restriction digests of genomic DNA and Southern hybridisation with Digoxigenin-labelled PCR-generated probes (procedure performed according to previously described protocols (28)) to detect differences in restriction patterns between the tested PAO sublines and the PAO1 reference sequence (31). Potential deletions were interrogated by combinatorial PCR taking the PAO1 sequence as a blueprint.

Pulsed field gel electrophoresis (PFGE).

Agarose plugs containing high molecular weight DNA were prepared by embedding 2×10^9 , 5×10^9 or 1×10^{10} bacterial cells / mL in low-melting-point agarose (Sigma), followed by proteinase K digestion at 56°C for 48 h. Restriction digests of these agarose plugs with *Swa*I, *Pac*I, *Dpn*I and *Spe*I were done as described previously (28), using the respective restriction buffer recommended by the manufacturer (New England Biolabs). PFGE of restricted DNA was performed in a CHEF DR III apparatus with 120° reorientation angle at 9.5°C with 0.5 x TBE (45 mM Tris, 45 mM boric acid, 1 mM EDTA, pH 8.3) as electrophoresis buffer. *Swa*I- and *Pac*I-digests were separated in 0.8 % agarose gels for 60 h at 2.5 V / cm (single ramp with linearly increasing pulse times from 500 to 2000 s). *Dpn*I- and *Spe*I-digests were separated by CHEF in 1.5 % agarose at 6 V / cm. Pulse times were linearly increased in three ramps from 8-50 s for 24 h, 12-25 s for 22 h and 1-14 s for 14 h. Restriction fragments were visualized with ethidium bromide.

Illumina Genome Analyzer Sequencing.

Illumina Genome Analyzer libraries were prepared following the manufacturer's instructions. Purified genomic DNA was fragmented by nebulization into fragments of less than 800 bp in size. The resulting 3' and 5' overhangs of the double-stranded DNA were converted into blunt ends and a terminal 'A' was then ligated to the 3' end. The modified DNA fragments were

ligated to adapters, purified and amplified by PCR according to the manufacturer's manual.

36 cycles of sequencing-by-synthesis were performed for each library in four lanes of a

flowcell with the Genome Analyzer. Illumina Genome Analyzer Pipeline Version 0.2

software was used to qualify reads passing default signal quality filters.

Each position in a sequence read is associated with a quality score Q_{solexa} reflecting the quality of the respective base call (depending on the measured signal during the sequencing process).

For a base X, the relation of its probability value $p(X)$ to the Q_{solexa} score is given by the formula

$$Q_{\text{solexa}} = 10 \log (p(X) / 1 - p(X)).$$

Therefore, a maximal Q_{solexa} value of 40 indicates a probability of 99.99 % for the called base, or an error probability of 0.01 %, respectively. Lower Q_{solexa} values typically appear at the end of reads. If no decisive base call could be made, the respective position in the sequence read was designated as 'N', (with Q_{solexa} set to -0.5 automatically).

The sequence reads were mapped to the PAO1-UW reference sequence (NC_002516) using Illumina's ELAND aligner (Efficient Large-Scale Alignment of Nucleotide Databases).

ELAND aligns a large number of short reads to reference sequences allowing up to two mismatches in the first 32 bases of a read.

Thereafter the frequency of coverage, base predictions and 'consensus qualities' were computed for each position of the reference sequence covered by aligned reads. The

'consensus quality' $Q_{\text{consensus}}$ for each position s in the reference sequence is defined by

$$Q_{\text{consensus}} (s) = Q_{\text{max}} / (Q (A) + Q (C) + Q (G) + Q (T)),$$

whereby $Q (A)$, $Q (C)$, $Q (G)$ and $Q (T)$ are the quality sums for each base and Q_{max} is the

maximal value thereof. Thus, a $Q_{\text{consensus}} (s)$ value of 1 indicates identical base calls for position s in all covering reads, while lower values are caused by heterogeneous results.

Assembly of sequence reads

Mapping the sequence reads obtained from the PAO-DSM and MPAO1 subline to the PAO1-UW reference sequence with ELAND allowed binning of the reads: 727 and 695 Mb of sequence were aligned to the reference ('in_ref bins'). 29.4 / 17.4 Mb of sequence ('not_in_ref bins') contained putatively novel DNA. Obviously incorrect reads with homooligomers > 13 bases in length (not present in the *P. aeruginosa* genome) or an 'N' – base call in at least three positions were excluded from the analysis.

The *de novo* assembler Velvet version 0.7.11 (35) was then used with a kmer size of 27 to generate contigs of novel DNA from the 'not_in_ref' sequences. The resulting sequence contigs were extended using an in-house program, MarkerCounter, to locate reads from the 'not_in_ref' bin. Assembly was completed by inclusion of reads from the 'in_ref bin'. Thus the novel DNA could be attributed to a new phage sequence, part of which shows 94% to 100% sequence identity with the PA0717 – PA0727 contig. The short read mapper Mosaik (The MarthLab, Boston College [<http://bioinformatics.bc.edu/marthlab/Mosaik>]) was used in tandem with Eagleview (12) to verify the variable positions in the duplicated sequence. The assembly and the presence of duplicated DNA were confirmed by combinatorial PCR. The exact size of the 1,006 bp deletion already detected in the initial assembly was determined by using MarkerCounter to identify reads covering the breakpoint.

SNPs and small deletions.

Analysis of the primary sequence data generated potential sequence variations (SNPs) defined as consensus bases differing from the published PAO1 reference. Forty-six SNP positions with varying coverage and confidence of base prediction were selected for validation by diagnostic restriction digests. Sequences encompassing the potential SNP position were amplified by PCR (lengths 163 – 292 bp). If the predicted nucleotide exchange resulted in gain or loss of a restriction site, the PCR product was digested with the respective enzyme. If no appropriate diagnostic enzyme could be found, artificial restriction sites were introduced

by one or two mismatches in a PCR primer (5). According to the outcome of the wet lab experiments the threshold for a reliable SNP in the primary sequence data was set to be an at least 10fold coverage and a $Q_{\text{consensus}}$ of 0.9 or more (see above).

PAO1-UW regions that were not covered by MPAO1- or PAO-DSM-derived sequence reads were candidates for deletions. These segments were queried for the number of deleted bases by direct blotting electrophoresis and subsequent chemoluminescence visualisation of combinatorial biotinylated PCR products (22).

Growth competition experiments.

Five mL LB broth was inoculated from glycerol stocks of the respective PAO1 sublines and grown overnight at 37°C with shaking. An aliquot was transferred into 20 mL LB adjusted to $OD_{578} = 0.1$ and cultivated for about 6 h until $OD_{578} = 2.5$ was reached. 2×10^{10} CFU per PAO1 subline were mixed and an aliquot thereof was added to 20 ml medium to give $OD_{578} = 0.1$. Table 2 describes the different culturing conditions of the competitive fitness experiments. At the respective endpoints, samples were taken and genomic DNA was prepared following standard procedures (see above). All cultures were performed in duplicate. The relative contents of PAO1 sublines were determined by semiquantitative PCR kinetics (15) utilizing subline-discriminating markers.

Murine airway infection model.

Bacteria were grown in LB broth overnight at 37 °C (230rpm), pelleted by centrifugation (5000 x g, 10min), washed twice with sterile phosphate buffered saline and the optical density of the bacterial suspension was adjusted by spectrophotometry at 578 nm. The intended number of CFU was extrapolated from a standard growth curve, and appropriate dilutions with sterile PBS were made to prepare the inoculum for the mice. To verify the correct dilution, an aliquot was serially diluted on LB agar plates. Ten to 12 week old female mice of

the inbred strain C3H/HeN (Charles River, Sulzfeld, Germany) were inoculated with 30 µl of this bacterial suspension containing 7.5×10^6 CFU of the different *P. aeruginosa* PAO1 sublines via view controlled intratracheal instillation. In case of 14-day infection experiments, the weight and rectal temperature of the mice were measured daily and their body condition was scored for the parameters vocalisation, piloerection, attitude, locomotion, breathing, curiosity, nasal secretion, grooming and dehydration (25).

During the experiments mice were maintained in microisolator cages with filter top lids at $21 \pm 2^\circ\text{C}$, 50 ± 5 % humidity and 12 h light-dark-cycle. They were supplied with autoclaved, acidulated water and fed *ad libitum* with autoclaved standard diet. Mice were sacrificed after 48 h for the evaluation of lung histology or the determination of CFU in homogenized organs (lungs, liver, spleen and brain).

Prior to the start of the experiments animals were acclimatized for at least seven days. All animal procedures were approved by the local District Governments and carried out according to the guidelines of the German law for the protection of animal life.

Database accession numbers.

The nucleotide sequence of the PAO-RGP42 island described in this paper has been deposited in the GenBank database (accession number GQ141978). The SNPs listed in Table 1 have been submitted to Database of Single Nucleotide Polymorphisms ‘dbSNP’ (submission SNP accession numbers ss138660650 – ss138660685).

Results

Large genomic differences between PAO1 sublines

PFGE of restriction digestions of PAO1-UW and PAO-DSM revealed discordant PacI and SmaI fragment patterns (Fig. 1A and Supplemental Material, Table S1), which visualize the

known inversion in PAO1-UW between the rDNA operons *rrnA* and *rrnB* (31). Please note that the *ter* locus is located in PAO1 sublines PAO-DSM and MPAO1 opposite to the origin of replication (Fig. 1B) as it has been observed in all other sequenced *P. aeruginosa* chromosomes (18, 21, 33). Hence, the first sequenced PAO1-UW strain with its asymmetric localisation of the terminus of replication has to be considered as an exception from the global genome architecture of *P. aeruginosa*. The inversion in PAO1-UW must have occurred after the PAO1 strain had been deposited by the original investigator in public strain collections. SpeI and DpnI restriction digestions yielded nearly indistinguishable fragment patterns of the PAO-UW and PAO-DSM chromosomes (Fig. 1C) because both SpeI and DpnI have recognition sites in the 23S rDNA sequence and hence do not visualise the inversion between the *rrn* operons. Closer inspection of the SpeI digest, however, showed a loss of a SpeI recognition site in PAO1UW (Fig. 1C). Fragments SpZ and SpAF of the initial PAO1 subline kept in the Holloway laboratory were replaced by fragment SpZ', which was smaller in size than the sum of the adjacent SpZ and SpAF fragments. Restriction mapping and sequencing of this region revealed a 12 kb insertion in the 3' end of a tRNA^{Met} (PA4673.1) in PAO-DSM and MPAO1 accompanied by a 1 kb deletion about 12 kb downstream of the insertion (Fig. 1D). The 12 kb insertion harbours the SpeI recognition site that is missing in the PAO1-UW chromosome.

The genomic island at PA4673.1 classified as RGP42 insertion (according to the nomenclature introduced (21)) encodes a putative protein phosphatase, two protein kinases and a prophage that is highly homologous to part of the Pfl-like phage inserted at the tRNA^{Gly} PA0729.1. The region spanning from ORF2 to ORF12 (Fig. 1E) is syntenic with PA0727 – PA0717 (RGP5) and shows 94% - 100% sequence identity between the individual homologs (Supplemental Material, Table S2). ORF1 codes for the phage-specific integrase that is identical in sequence with integrases in other *P. aeruginosa* strains like 2192 that targeted the same tRNA^{Met} gene. ORFs 14-16 are specific for the genomic island. ORF14 codes for a

protein phosphatase and ORFs15 and 16 each for a protein kinase. A homologous operon with the same organization of ORFs has to date only been observed in the genome of the Mono Lake delta-proteobacterium MLMS-1 (8). Due to the insertion, 46 bp of the 3' terminal sequence of the tRNA^{Met} gene and 36 bp downstream were duplicated and flank the insertion at both sites. The presence of flanking tRNA 3' ends is typical for genomic islands in *P. aeruginosa* (14).

PAO-DSM and MPAO1 lack the PAO1-UW sequence from 5,253,694 to 5,254,699. This in-frame deletion disrupts the conserved hypotheticals PA4684 (1299 bp) and PA4685 (696 bp) and yields a chimeric ORF PA4684' with the initial 936 bp of PA4684 and the terminal 42 bp of PA4685. The PAO1 genomic gene arrays showed above average cDNA hybridization signals for the PA4684' fusion gene of PAO-DSM during growth in LB broth in late exponential phase in presence and absence of various exogenous stresses indicating that PA4684' is transcribed. Full-length PA4684 and PA4685 genes are present in all other yet completely sequenced *P. aeruginosa* demonstrating that the in-frame deletion is a unique feature of a fraction of PAO1 sublines.

SNPs and small deletions

Genome sequencing of the chromosomes of PAO1 sublines MPAO1 and PAO-DSM by Genome Analyzer sequencing-by-synthesis technology uncovered 39 validated single nucleotide deviations (SNPs) from the PAO-UW database entries (Table 1, Figure 2), six of which are pairs of adjacent nucleotides. Twenty-two SNPs are transversions and 17 SNPs are transitions. Nine SNPs are located in intergenic regions of which SNP 5,036,891 resides within one of the non-coding RNA loci described so far (3, 20, 30). Of the 30 SNPs in coding regions, 13 are synonymous substitutions, 16 cause an amino acid exchange and one changes an arginine into a stop codon. Several amino acid exchanges are non-conservative and should influence structure and/or function of the protein. Examples include the change from

phenylalanine to serine in the nitrate reductase NapA involved in anaerobiosis, or from aspartate to tyrosine in the cell division protein FtsZ (Table 1).

The initial analysis of the sequencing data pointed to the potential existence of 140 SNPs in PAO-DSM and 153 SNPs in MPAO1 compared to the published PAO1 sequence. A minimal 10fold coverage of a nucleotide position and a quality value of 0.9 or more for a nucleotide substitution (see Methods section) were considered to indicate a SNP and not a sequencing artefact. Thirty-seven SNPs in PAO-DSM and 22 SNPs in MPAO1 passed this threshold criterion. Thirteen of 14 candidate SNPs that fulfilled the quality criteria were verified to be nucleotide substitutions by informative restriction digestions of PCR products of PAO-UW1, PAO-DSM and MPAO1 DNA templates (Supplemental Material, Table S3). In contrast, a nucleotide substitution was experimentally proven for only one of 32 tested candidate SNPs that did not pass the threshold. Hence, the rigorous threshold criteria applied to the primary sequencing data were appropriate to identify more than 95% of the SNPs with an acceptable rate of false positives of less than 10%.

Table 1 lists the nucleotide position of 39 validated SNPs, one of which was found only in MPAO1 and 16 of which in PAO-DSM only. Nine strain-specific synonymous SNPs are located in ORFs PA0725 and PA0726 of the RGP5 prophage region that is duplicated in PAO-DSM and MPAO1, but not in PAO1-UW (see above). Five non-synonymous and one synonymous nucleotide substitutions were also detected by diagnostic restriction digestions of PCR products in the PAO-UW strain maintained in our laboratory. Figure 2 shows the genome map positions of the SNPs.

The Genome Analyzer sequencing identified 17 regions that contained putative insertions or deletions. Subsequent combinatorial PCR and direct blotting electrophoresis verified four small deletions (Figure 2). First, the PAO1-UW subline maintained in our laboratory showed intragenic deletions in PA1695 and PA3760 compared to the published PAO1-UW sequence. Second, the PAO-DSM subline lacked two of the 18 heptanucleotide repeats in the terminal

region of the RGP5 prophage downstream of PA0715 and carried a 3-bp deletion within PA0406. The latter deletion (447,948 – 447,950) affected two codons but resulted in the loss of a glutamine codon only while the neighboured lysine codon is restored.

Comparative fitness of PAO1 sublines in vitro and in vivo

Next, we studied the global impact of the sequence diversity among sublines on competitive growth under various in vitro conditions (Table 2) and on virulence in an acute murine airway infection model (24).

If the three PAO1 sublines were cultured in parallel in LB in separate flasks, the growth behaviour and doubling times of the three strains were identical within experimental error. However, when subsequently equal numbers of CFU of the three PAO1 sublines were mixed and then inoculated into fresh medium, subline PAO-DSM was found to be completely eliminated. The DSM-specific SNPs at positions 4,769,678 and 2,807,853 were never detected in up to 35 cycles of diagnostic PCR of the genomic DNA of the bacterial mixture grown under seven different experimental conditions (Table 2). The positive control of mixtures of equal amounts of genomic DNA of the three PAO1 sublines grown in separate flasks showed agarose-gel separated PCR products in expected stoichiometric ratios (data not shown). PAO-DSM DNA was absent at all time points including the first sample processed within the first 10 minutes after mixing. Hence we hypothesized that PAO-DSM was killed by MPAO1 and/or PAO-UW. Exposure of LB agar-embedded PAO-DSM to supernatants of either MPAO1 or PAO-UW cultures killed the PAO-DSM bacteria indicating that MPAO1 and PAO-UW had lysed the susceptible PAO-DSM host and/or cleaved its DNA. In all other combinations the PAO1 host bacteria remained viable.

Thus, only MPAO1 and PAO-UW grew together. Bacterial mixtures were harvested in either exponential or stationary phase and their genomic DNA was analysed by quantitative PCR kinetics for the presence of the deletion in MPAO1 comprising nucleotides 5,253,694 to

5,254,699 (see above). According to the PCR data both strains were present in similar amounts under most investigated conditions. However, PAO1-UW made up 80% and more of the bacterial population in stationary phase when growing in LB broth at 42°C and in (M9 + succinate) minimal medium at 37°C. PAO1-UW had outcompeted MPAO1 during heat stress or limited nutrient supply indicating that the two sublines display differential fitness under distinct stress conditions.

Virulence of the PAO1 sublines was tested by intratracheal instillation of exponentially growing bacteria into the lungs of C3H/HeN mice (24). A dose of 10^6 CFU caused a moderate acute pneumonia by day 2 and led to death of about half of the animals by day 5 post infection (Figure 3). A tenfold higher dose of 10^7 CFU had a similar outcome for infections with MPAO1, but led to a profound purulent pneumonia and subsequent death of all animals between day 2 and 3 for infections with PAO-DSM or PAO1-UW. Thus, a differential pathogenic potential of PAO1 sublines became evident at fatal doses.

Discussion

The PAO1 strain remains the major reference strain for research on *P. aeruginosa*. Physical mapping and whole genome re-sequencing of PAO1 sublines uncovered an extended genomic diversity among the two strains that were the basis for the initial sequencing project and the first genome-wide mutant library. Moreover, progeny of the strain PAO1 that had been identical in its physical map with the original strain stored in Dr. Holloway's laboratory (11) turned out to harbour numerous deviations from the published PAO1 sequence. The major differences of sublines MPAO1 and PAO-DSM from the genomic sequence of the strain PAO-UW (31) are a synteny of genes identical to that of all other sequenced *P. aeruginosa* strains (18, 21, 33), a duplication of a RGP prophage carrying a distinct integrase and putative protein serine/threonine phosphotransferases of unknown function as well as numerous SNPs

and short indels. These variations, that make up about 0.2% of the whole genome sequence, were found to be associated with a strong impact on fitness and virulence. Most unexpectedly the MPAO1 and PAO-UW sublines killed PAO-DSM immediately when strains were pooled together to one bacterial population. Intraspecies killing in *P. aeruginosa* is typically mediated by either phages or pyocins. No sequence diversity among the PAO1 sublines was detected in pyocin-encoding genes. In contrast, multiple SNPs were identified in PAO-DSM in two ORFs of the duplicated prophage region. Although none of these nucleotide substitutions in PA0725 and PA0726 led to an amino acid change, it is tempting to assume that these SNPs affect the ability of PAO-DSM to survive in the presence of progeny of the same ancestor.

The reported SNPs and small indels represent a conservative lower estimate of the actual sequence variation. The chosen threshold for the number and quality of sequence reads selected those sequence variations identified by the Illumina Genome Analyzer that represent real SNPs with more than 95% probability (see Results). According to our wet lab data up to ten further SNPs may exist in the genome that were not detected with sufficient specificity to pass the quality criterion.

These estimates, however, only apply to homogeneous sequence variations in all cells of a PAO1 subline. One major objective of novel next-generation sequencing technologies is the identification of sequence variations that are present in a minority of molecules in a DNA sample. Ultra-deep sequencing of loci that are diagnostic for minimum residual disease in human cancer would be a typical application (23). Sequencing of microbial genomes by the Genome Analyzer also pinpoints positions with more than one nucleotide in the different sequence reads. If we exclude the technical artefacts that are typically seen at the end of the short reads, the read counts of the various sequence variants represent the proportions of distinct sequence isoforms in the bacterial population. An instructive example within our project is the known hotspot of mutation in the *P. aeruginosa* genome, the gene *lasR*

encoding the major transcriptional regulator for quorum sensing in *P. aeruginosa* (17, 29).

The sequenced DNA sample of the PAO-DSM strain revealed a G-T transversion in 9 of 79 sequence reads resulting in a substitution of arginine in leucine in about 10% of the population. This single amino acid substitution leads to pleiotropic changes of phenotype, including defective LasR function, overproduction of pyocyanin, impaired utilization of adenosine as carbon source and higher resistance to cell lysis in stationary phase (6). The laboratory of one of the authors (SH) sequenced individual clones selected from PAO-DSM stock cultures of two laboratories (Miguel Cámara, personal communication). The two *lasR* sequences were present in both stocks. This case nicely illustrates the on-going microevolution of the PAO1 reference strain that remains unnoticed by researchers until some unexpected changes in phenotype emerge.

The two variants of the PAO-UW strain kept in our laboratory further substantiate the genotypic diversification of PAO1 during its worldwide distribution. The strain received in 1999 from the genome sequencing team showed the diagnostic features of the large inversion and the small 12 kb deletion of the island RGP42 (Fig. 1). On the other hand, the genome of the strain received in 2005 and also designated as ‘PAO1-UW’ harboured the complete PA4684-PA4685 genes like the sequenced PAO1-UW, but also, like PAO-DSM or MPAO1, contained RGP42. This ‘PAO1-UW’ strain moreover shared six sequence deviations from the published PAO1 sequence with PAO-DSM and MPAO1 (Table 1). The variable presence of RGP42 in PAO1-UW and ‘PAO1-UW’ is consistent with our notion during physical mapping that the diagnostic *SpeI* fragment exhibited a variable intensity of the ethidium bromide stain in separate PFGE runs implicating that this prophage region is prone to deletion. Besides this inherent genomic plasticity of the PAO1 strains the individual handling of the strain may drive microevolution. A member of our laboratory, for example, unintentionally generated a hypermutator variant of PAO-DSM by continuous subculturing. In summary, intrinsic

biological features of the strain and laboratory procedures have caused a broad spectrum of PAO1 genotypes during the last 50 years.

Some nucleotide substitutions affected genes such as *fusAI* (PA4266) and *ftsZ* (PA4407) that encode basal functions of the bacterial cell. *FusAI* encodes the elongation factor G that during translation is involved in the movement of peptidyl-tRNA from the A site to the P site and in the disassembly of the post-termination ribosomal complex (7, 34). The PAO1 variants carry either a threonine or isoleucine at position 493 of the 706 aa long protein. Position 493 is located at the beginning of the functional domain 4 with a close to 100% conserved sequence of the next ten amino acids in N-terminal direction and a flexible region with numerous exchanges among the next ten amino acids in C-terminal direction. Of the 70 most related EF-G proteins in the NCBI database, threonine is the dominant amino acid at position 493, whilst isoleucine is not present. Consequently we hypothesize that the amino acid exchange is functionally not neutral. A formally more drastic substitution of an aspartate to tyrosine was detected in the cell division protein FtsZ (PA4407) at position 379 close to its C-terminus. The tubulin-like FtsZ forms long dynamic spirals that are reorganized into a ring structure at the onset of division (4). The amino acid substitution is located in the least conserved region of FtsZ and thus - despite the pronounced impact on the local environment – should not influence the global function of this core element of cell division. A further non-conservative phenylalanine to serine substitution affected the nitrate reductase NapA (PA1174). Nitrate reductases enable *P. aeruginosa* to grow under microaerophilic or anaerobic conditions by denitrification. However, it is the membrane-bound nitrate reductase NarG, but not the periplasmic enzyme NapA, that is essential to sustain anaerobic growth (26). Although NapA is more strongly expressed during anaerobiosis (own unpublished transcriptome data), the sequence variations in the NapA protein should not significantly influence the global fitness of the PAO1 strains. In summary, these three amino acid substitutions can be classified as modulators or benign sequence variants.

Besides the loss-of-function mutation in the *lasR* gene of a subpopulation of PAO-DSM bacteria, a further non-conservative substitution of alanine to proline was identified in MexT, a further hotspot of mutation in the *P. aeruginosa* genome (29). MexT activates the expression of the MexEF-OprN efflux pump and represses the expression of the outer membrane porin protein, OprD (16). Both porin and efflux pump exhibit broad specificity, and hence functionally non-equivalent MexT isoforms could promote pleiotropic effects on the transport of small molecules across the cell membrane and thus on the global transport profile of the cell.

The sequence variants in PA4498 either encoded a 405 amino acid long protein or a truncated version of 34 amino acids (Table 1). PA4498 mRNA transcript was very lowly expressed in six tested *P. aeruginosa* strains. PAO-DSM, however, which harbours the premature termination codon, displayed more than 10-fold higher mRNA transcript levels indicating that the full-length protein is needed. Correspondingly, we classify this substitution of an arginine to a stop codon as a deleterious mutation in PAO-DSM. Thus, the latter two cases of sequence diversity in *mexT* and PA4498 could contribute to the diversification of phenotype among the PAO1 sublines.

In conclusion, the maintenance and propagation of *P. aeruginosa* PAO1 in laboratories throughout the world have entertained an on-going microevolution of genotype and phenotype that jeopardizes the reproducibility of research. PAO1 is still the major reference strain for the *P. aeruginosa* field and hence – with the data of this study at hand – we now envisage a similar complexity as has been recognized in the past for *E. coli* K12, the major workhorse for biochemists, molecular biologists and bacterial geneticists. Standardization of genotype by accredited strain collections is advisable. High-throughput genome re-sequencing will resolve more cases and could become a proper quality control for strain collections.

Acknowledgments

This work was partially supported by grants provided by the Deutsche Forschungsgemeinschaft (SFB 587, project A9; GRK 653/3). CFD is a member of the DFG-supported IRTG 'Pseudomonas: Pathogenicity and Biotechnology'.

The authors thank Miguel Cámara and Jean Dubern for the thorough analysis of the *lasR* locus in PAO-DSM stock cultures and Silke Hedtfeld for the verification of small indels by direct blotting electrophoresis. Supply of PAO1 sublines by Alice L. Erwin and Thomas Rasmussen is gratefully acknowledged.

References

1. **Ausubel F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidmann, J. A. Smith, and K. Struhl** (ed). 1994. Current protocols in molecular biology. Wiley, New York, NY
2. **Fox A., D. Haas, C. Reimmann, S. Heeb, A. Filloux, and R. Voulhoux.** 2008.
Emergence of secretion-defective sublines of *Pseudomonas aeruginosa* PAO1 resulting from spontaneous mutations in the *vfr* global regulatory gene. Appl Environ Microbiol. **74**:1902-1908.
3. **González N., S. Heeb, C. Valverde, E. Kay, C. Reimmann, T. Junier, and D. Haas.** 2008. Genome-wide search reveals a novel GacA-regulated small RNA in *Pseudomonas* species. BMC Genomics. **9**:167. doi:10.1186/1471-2164-9-167.
4. **Graumann P. L.** 2007. Cytoskeletal elements in bacteria. Annu Rev Microbiol. **61**:589-618.
5. **Grimes E., M. Koob, and W. Szybalski.** 1990. Achilles' heel cleavage: creation of rare restriction sites in lambda phage genomes and evaluation of additional operators, repressors and restriction/modification systems. Gene. **90**:1-7.
6. **Heurlier K., V. Dénervaud, M. Haenni, L. Guy, V. Krishnapillai, and D. Haas.** 2005. Quorum-sensing-negative (*lasR*) mutants of *Pseudomonas aeruginosa* avoid cell lysis and death. J Bacteriol. **187**:4875-4883.
7. **Hirokawa G., N. Iwakura, A. Kaji, and H. Kaji.** 2008. The role of GTP in transient splitting of 70S ribosomes by RRF (ribosome recycling factor) and EF-G (elongation factor G). Nucleic Acids Res. **36**:6676-6687.
8. **Hoeft S.E., T. R. Kulp, J. F. Stolz, J. T. Hollibaugh, and R. S. Oremland.** 2004. Dissimilatory arsenate reduction with sulfide as electron donor: experiments with mono lake water and Isolation of strain MLMS-1, a chemoautotrophic arsenate respirer. Appl Environ Microbiol. **70**:2741-2747.

9. **Holloway B.W.** 1955. Genetic recombination in *Pseudomonas aeruginosa*. J Gen Microbiol. **13**:572-581.
10. **Holloway B.W., V. Krishnapillai, and A. F. Morgan AF.** 1979. Chromosomal genetics of *Pseudomonas*. Microbiol Rev. **43**:73-102.
11. **Holloway B.W., U. Römling, and B. Tümmler.** 1994. Genomic mapping of *Pseudomonas aeruginosa* PAO. Microbiology. **140**:2907-2929.
12. **Huang W., and G. Marth.** 2008. EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res. **18**:1538-1543.
13. **Jacobs M. A., A. Alwood, I. Thaipisuttikul, D. Spencer, E. Haugen, S. Ernst, O. Will, R. Kaul, C. Raymond, R. Levy, L. Chun-Rong, D. Guenther, D. Bovee, M. V. Olson, and C. Manoil.** 2003. Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. Proc Natl Acad Sci USA. **100**:14339-14344.
14. **Klockgether J., O. N. Reva, and B. Tümmler.** 2006. Spread of genomic islands between clinical and environmental strains, p. 187-200. In N. A. Logan, H. M. Lapin-Scott, P. C. F. Oyston (ed.), Prokaryotic diversity: mechanisms and significance. Cambridge University Press, Cambridge, UK.
15. **Klockgether J., D. Würdemann, L. Wiehlmann, and B. Tümmler.** 2008. Transcript profiling of the *Pseudomonas aeruginosa* genomic islands PAGI-2 and pKLC102. Microbiology. **154**:1599-1604.
16. **Köhler T., S. F. Epp, L. K. Curty, and J. C. Pechère.** 1999. Characterization of MexT, the regulator of the MexE-MexF-OprN multidrug efflux system of *Pseudomonas aeruginosa*. J Bacteriol. **181**:6300-6305.
17. **Köhler T., A. Buckling, and C. van Delden.** 2009. Cooperation and virulence of clinical *Pseudomonas aeruginosa* populations. Proc Natl Acad Sci U S A. **106**:6339-6344.
18. **Lee D. G., J. M. Urbach, G. Wu, N. T. Liberati, R. L. Feinbaum, S. Miyata, L. T. Diggins, J. He, M. Saucier, E. Déziel, L. Friedman, L. Li, G. Grills, K. Montgomery,**

- R. Kucherlapati, L. G. Rahme, and F. M. Ausubel.** 2006. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol.* **7**:R90. doi: 10.1186/gb-2006-7-10-r90.
19. **Lewenza S., R. K. Falsafi, G. Winsor, W. J. Gooderham, J. B. McPhee, F. S. Brinkman, and R. E. Hancock.** 2005. Construction of a mini-Tn5-*luxCDABE* mutant library in *Pseudomonas aeruginosa* PAO1: a tool for identifying differentially regulated genes. *Genome Res.* **15**:583-589.
20. **Livny J., A. Brencic, S. Lory, and M. K. Waldor.** 2006. Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res.* **34**:3484-3493.
21. **Mathee K., G. Narasimhan, C. Valdes, X. Qiu, J. M. Matewish, M. Koehrsen, A. Rokas, C. N. Yandava, R. Engels, E. Zeng, R. Olavarietta, M. Doud, R. S. Smith, P. Montgomery, J. R. White, P. A. Godfrey, C. Kodira, B. Birren, J. E. Galagan, and S. Lory.** 2008. Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci U S A.* **105**:3100-3105.
22. **Mekus F., U. Laabs, H. Veeze, and B. Tümmler.** 2003. Genes in the vicinity of CFTR modulate the cystic fibrosis phenotype in highly concordant or discordant F508del homozygous sib pairs. *Hum Genet.* **112**:1-11.
23. **Morozova O., and M. A. Marra.** 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics.* **92**:255-264.
24. **Munder A., S. Krusch, T. Tschernig, M. Dorsch, A. Lührmann, M. van Griensven, B. Tümmler, S. Weiss, and H. J. Hedrich.** 2002. Pulmonary microbial infection in mice: comparison of different application methods and correlation of bacterial numbers and histopathology. *Exp Toxicol Pathol.* **54**:127-133.
25. **Munder A., A. Zelmer, A. Schmiedl, K. E. Dittmar, M. Rohde, M. Dorsch, K. Otto, J. H. Hedrich, B. Tümmler, S. Weiss, and T. Tschernig.** 2005. Murine pulmonary

infection with *Listeria monocytogenes*: differential susceptibility of BALB/c, C57BL/6 and DBA/2 mice. *Microbes Infect.* **7**:600-11.

26. **Palmer K.L., S. A. Brown, and M. Whiteley.** 2007. Membrane-bound nitrate reductase is required for anaerobic growth in cystic fibrosis sputum. *J Bacteriol.* **189**:4449-4455.
27. **Römling U., D. Grothues, W. Bautsch, and B. Tümmler.** 1989. A physical genome map of *Pseudomonas aeruginosa* PAO. *EMBO J.* **8**:4081-4089.
28. **Römling U., T. Heuer, and B. Tümmler.** 1994. Bacterial genome analysis by pulsed field gel electrophoresis techniques. *Adv Electrophoresis.* **7**:353-406.
29. **Smith E. E., D. G. Buckley, Z. Wu, C. Saenphimmachak, L. R. Hoffman, D. A. D'Argenio, S. I. Miller, B. W. Ramsey, D. P. Speert, S. M. Moskowitz, J. L. Burns, R. Kaul, and M. V. Olson.** 2006. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A.* **103**: 8487-8492.
30. **Sonnleitner E., T. Sorger-Domenigg, M. J. Madej, S. Findeiss, J. Hackermüller, A. Hüttenhofer, P. F. Stadler, U. Bläsi, and I. Moll.** 2008. Detection of small RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatic tools. *Microbiology.* **154**: 3175-3187.
31. **Stover C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrenner, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson.** 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature.* **406**:959-964.
32. **Winsor G. L., T. Van Rossum, R. Lo, B. Khaira, M. D. Whiteside, R. E. Hancock, and F. S. Brinkman.** 2009. *Pseudomonas* Genome Database: facilitating user-friendly,

comprehensive comparisons of microbial genomes. *Nucleic Acids Res.* **37(Database issue)**:D483-488.

33. **Winstanley C., M. G. Langille, J. L. Fothergill, I. Kukavica-Ibrulj, C. Paradis-Bleau, F. Sanschagrín, N. R. Thomson, G. L. Winsor, M. A. Quail, N. Lennard, A. Bignell, L. Clarke, K. Seeger, D. Saunders, D. Harris, J. Parkhill, R. E. Hancock, F. S. Brinkman, and R. C. Levesque.** 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.
34. **Yu H., Y. L. Chan, and I. G. Wool.** 2009. The identification of the determinants of the cyclic, sequential binding of elongation factors tu and g to the ribosome. *J Mol Biol.* **386**:802-813.
35. **Zerbino D. R., and E. Birney.** 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**:821–829.

Figure Legends

Figure 1. Genomic differences in PAO sublines.

A. PFGE of *Swa*I- and *Pac*I-digested DNA of PAO sublines PAO-DSM (a) and PAO1-UW (b). Restriction fragments that are altered due to the chromosomal inversion are labelled.

B. Circular maps of PAO sublines PAO-DSM and PAO1-UW showing the localization of the *ori* and *ter* regions, *Swa*I (S) and *Pac*I (P) restriction sites and ribosomal operons. The part of the genome shown in grey is inverted in PAO1-UW in contrast to PAO-DSM and other sublines. The inversion breakpoints (*inv*) are located within *rrnA* and *rrnB*. Genome coordinates are given in accordance with the published PAO1-UW sequence (31). The black dot indicates the genomic region with a complex insertion-deletion event shown in C and D.

C. PFGE of *Spe*I- and *Dpn*I-digested DNA of PAO sublines. *Dpn*I-restriction patterns were not distinguishable for PAO-DSM (a, left part) and PAO1-UW (b, right part). However, *Spe*I patterns show fragments SpZ and SpAF of PAO-DSM replaced by SpZ' in PAO1-UW.

D. Physical map of the genomic region with altered *Spe*I pattern, based on the PAO1-UW reference sequence. 1006 bp of the reference sequence (white box) in proximity to *rrnC* are missing in sublines PAO-DSM and MPAO1. Instead, both these sublines harbour 12,066 bp of additional DNA (including a further *Spe*I restriction site) located next to the indicated tRNA^{Met} PA4673.1.

E. Gene map of the 12,066 bp insertion. Homologs of phage-like ORFs PA0717 – PA0727 in PAO1-UW are shown with a mesh-like pattern, genes encoding typical phage proteins are coloured in grey (integrase: *int*, coat proteins: *coaAB*). Black arrows indicate ORFs encoding serine/threonine phosphotransferases missing in PAO1-UW.

Figure 2. Genomic distribution of SNPs and indels.

Positions of SNPs (bars) in *P. aeruginosa* sublines MPAO1 (inside the circle) and PAO-DSM (outside) and of indels (triangles) in PAO-DSM (1, 2) and the tested PAO1-UW subline (3, 4). a, nine substitutions in PAO-DSM within 725 bp; b, one substitution in both sublines and another one in PAO-DSM 130 bp upstream; c,d, two substitutions in both sublines at adjacent positions; e, two substitutions in both sublines at adjacent positions and one substitution in PAO-DSM 1400 bp downstream.

Figure 3. Murine airway infection.

Pathohistological findings in C3H/HeN murine lungs 48 h after intratracheal infection with 10^6 CFU or 10^7 CFU of *P. aeruginosa* PAO1 sublines or vehicle control (Hematoxylin-eosin staining; original magnification x 100, scale bars 100 μ m).

Upper row: Infection with 10^6 CFU of all three PAO1 strains has induced a moderate acute fibrinogenic pneumonia, mostly confined to the alveoli. Bronchi are almost free of inflammatory cells. Vehicle control shows normal lung parenchyma after PBS application. Middle row: Animals infected with 10^7 CFU of PAO1-UW and PAO-DSM show profound purulent pneumonia with numerous foci of inflammation (infiltration of neutrophils and alveolar macrophages) up to tissue necrosis. In contrast strain MPAO1 has only generated moderate inflammation comparable to the infection with 10^6 CFU. The right insert depicts the colony forming units recovered from murine organs (lu, lung; spl, spleen; li, liver; br, brain) 48 h after intratracheal instillation of 10^6 CFU (closed symbols) or of 10^7 CFU (open symbols) of PAO-DSM (squares), MPAO1 (circles) or PAO1-UW (triangles).

Lower row: (left) Survival of mice whose airways had been inoculated with 10^6 CFU (closed symbols) or of 10^7 CFU (open symbols) of PAO-DSM (squares), MPAO1 (circles) or PAO1-UW (triangles). (right) Body condition score at day 1 after infection (untroubled (u); slightly troubled (▼); moderately troubled (▼▼); profoundly troubled (▼▼▼); moribund; death).

Table 1: List of SNPs in PAO-DSM and MPAO1**SNPs in both strains:**

position	dbSNP acc. no.¹	locus	encoded product	nt	aa
169284	ss138660650	interg. (PA0148-49)	-	G → C	
183697	ss138660651	PA0159	prob. transcriptional regulator	T → G	C → W
413850	ss138660652	interg. (PA0369-70)	-	T → C	
1467483	ss138660654	interg. (PA1352-53)	-	C → G	
2239547	ss138660657	interg. (PA2046-47)	-	T → G	
4212201	ss138660660	PA3760	prob. phosphotransferase	A → G	H → R
4344266	ss138660661	PA3877	nitrite extrusion protein 1 NarK	A → G	syn.
4448855	ss138660662	interg. (PA3969-70)	-	C → G	
4448856	ss138660662	interg. (PA3969-70)	-	G → C	
4869855	ss138660663	PA4341	prob. transcriptional regulator	T → G	E → N
4924552	ss138660664	PA4394	hypothetical protein	C → G	V → L
4924553	ss138660664	PA4394	hypothetical protein	G → C	syn.
5036891 ²	ss138660665	interg. (PA4499-500)	-	A → C	
5743461	ss138660666	PA5100	urocanate hydratase HutU	C → G	T → S
5743462	ss138660666	PA5100	urocanate hydratase HutU	G → C	T → S
6098781	ss138660668	PA5418	sarcosine oxidase subunit SoxA	G → C	syn.

SNPs in PAO-DSM only:

position	dbSNP acc. no.¹	locus	encoded product	nt	aa
447942	ss138660670	PA0406	TonB-like periplasmic protein	T → C	R → G
792696 ³	ss138660671	PA0725	phage Pfl-like hypoth. protein	C → T	syn.
792834 ³	ss138660672	PA0725	phage Pfl-like hypoth. protein	C → T	syn.
792882 ³	ss138660673	PA0725	phage Pfl-like hypoth. protein	T → G	syn.
793143 ³	ss138660674	PA0726	phage Pfl-like hypoth. protein	T → C	syn.
793242 ³	ss138660675	PA0726	phage Pfl-like hypoth. protein	T → G	syn.

793284 ³	ss138660676	PA0726	phage Pfl-like hypoth. protein	A → G	syn.
793347 ³	ss138660677	PA0726	phage Pfl-like hypoth. protein	G → A	syn.
793389 ³	ss138660678	PA0726	phage Pfl-like hypoth. protein	T → C	syn.
793410 ³	ss138660679	PA0726	phage Pfl-like hypoth. protein	T → C	syn.
2807853	ss138660680	PA2492	transcriptional regulator MexT	G → C	A → P
4769678	ss138660681	PA4266	elongation factor G (FusA1)	G → A	T → I
4939350	ss138660682	PA4407	cell division protein FtsZ	C → A	D → Y
5033102 ⁴	ss138660683	interg. (PA4496-97)	-	T → C	
5035975	ss138660684	PA4498	prob. metallopeptidase	G → A	R → stop ⁵
5744862	ss138660685	interg. (PA5100-01)	-	T → C	

SNP in MPAO1 only:

position	dbSNP acc. no. ¹	locus	encoded product	nt	aa
1275767 ⁶	ss138660653	PA1174	periplasmic nitrate reductase NapA	A → G	F → S

Base exchanges in both strains, also detected for PAO1-UW (Lyngby):

position	dbSNP acc. no. ¹	locus	encoded product	nt	aa
1589438	ss138660655	PA1459	chemotaxis-specific methylesterase	G → C	G → A
1835045	ss138660656	PA1685	enolase-phosphatase E1 (MasA)	G → C	S → T
2669175	ss138660658	PA2400	prob. non-ribosomal peptide synthetase PvdJ	G → C	P → A
2807982	ss138660659	PA2492	transcriptional regulator MexT	T → A	F → I
6079222	ss138660667	PA5399	prob. ferredoxin	A → G	syn.
6115455	ss138660669	PA5434	tryptophane permease (Mtr)	T → G	K → N

SNPs / exchanges printed in bold have been validated by diagnostic restriction digest

¹: submitted SNP accession numbers

²: occurs in predicted small RNA ((30), genome position 5,036,879 – 5,036,995; multiz – predicted locus 176, window 743)

³: located within ORF of RGP5 prophage sequence

⁴: base at this position uncertain in MPAO1

⁵: would cause truncated protein of only 34 aa (in comparison to 405 aa in PAO1-UW)

⁶: base at this position uncertain in PAO-DSM

Table 2. Conditions for growth competition experiments of PAO1 sublines

medium	temperature	shaking	an/aerobiosis	sampling times
LB broth	37°C	yes	aerobic	0.2 h; 0.5 h; 1 h; 2 h; 5 h; 16 h
M9 + succinate	37°C	yes	aerobic	16 h; 22 h
LB broth	37°C	no	microaerophilic	16 h; 40 h
LB broth	42°C	no	microaerophilic	16 h; 40 h
LB broth	4 °C	yes	aerobic	88 h; 162 h
LB broth	37 °C	yes	anaerobic	24 h; 48 h
LB broth	22 °C	yes	aerobic	5 h; 16 h

Figure 1

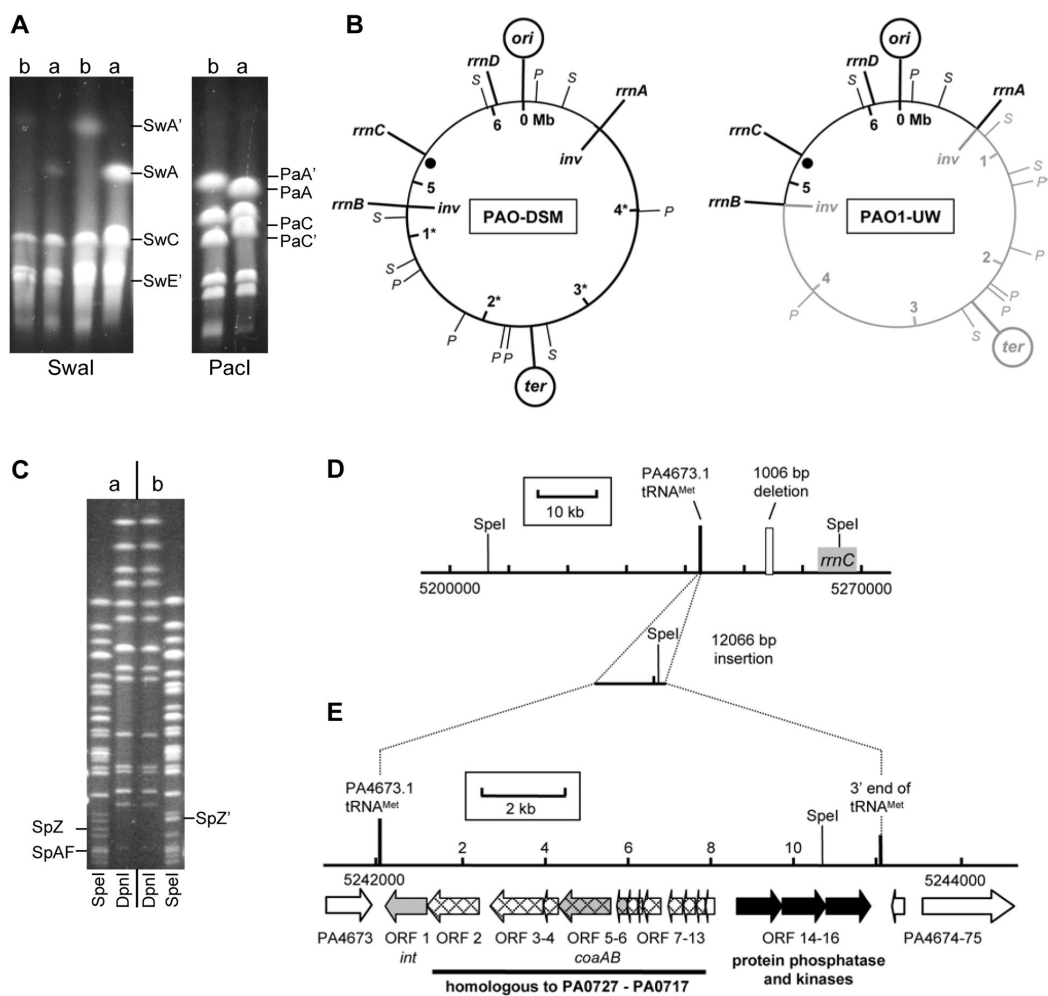


Figure 2

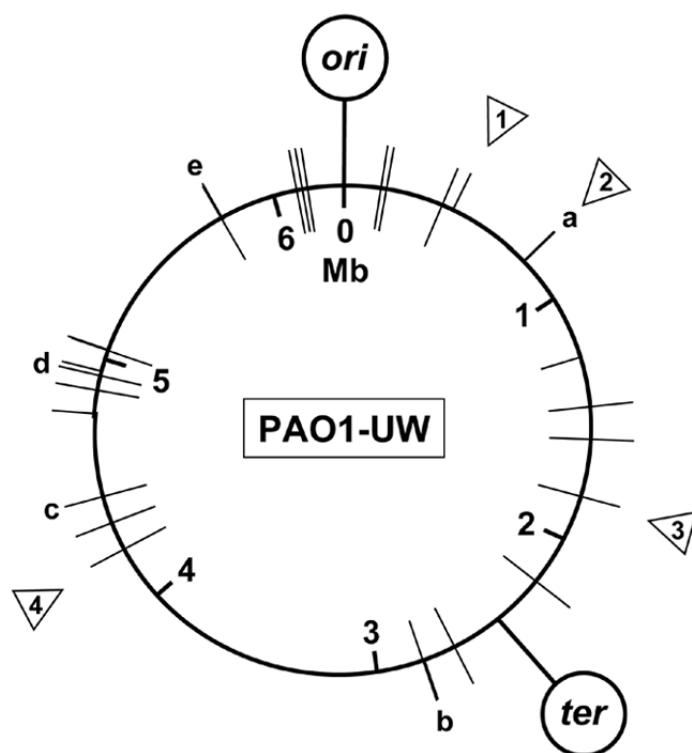
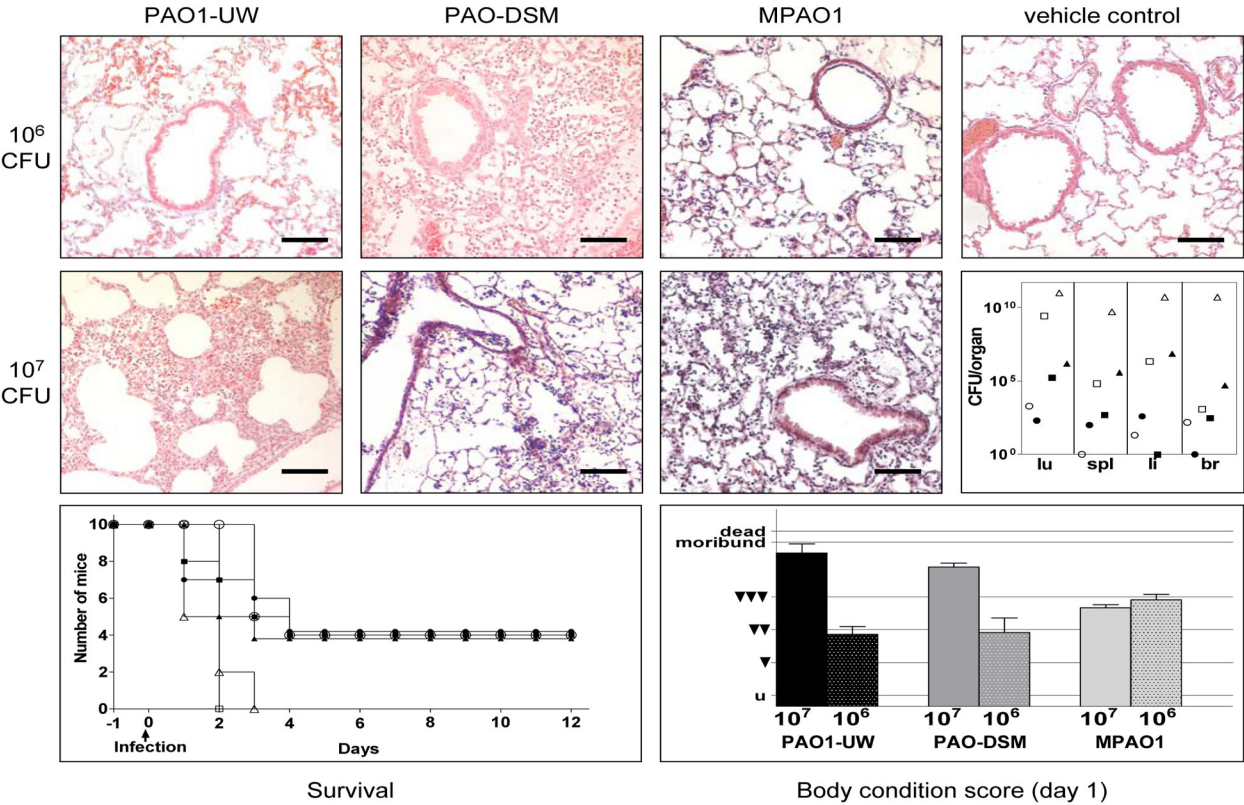


Figure 3



5. Metatie - a fast, accurate and adaptable metagenomic classifier for ultra-short reads

5.1 Background

At the time of writing metagenomics is an exciting topic. Sequencing technologies are plummeting in cost and providing many laboratories access to the "uncultured majority" of microbial organisms (Rappé and Giovannoni 2003, Hugenholtz and Tyson 2008). Bioinformaticians have already produced a range of programs for analysis, although these are "still in their infancy" (McHardy et al. 2007). Algorithm speed remains a problem (D. Huson, *pers. comm.*), with the most accurate solutions still around the same speed as Blast (Brady and Salzberg 2009). This is however far too slow for the metagenomic analysis of short reads in appropriate timeframes. Faster analytical approaches will start to arrive soon to meet this research need, such as an implementation of theSeed adapted to metagenomics (Aziz et al. 2008).

The random sequencing of reads used in the methodology described above has distinct advantages over the traditional use of rRNA (Manichanh et al. 2008, Biers et al. 2009). Ribosomal RNA has been used extensively for the best part of two decades in exploration of microbial communities. This marker gene can be assessed in environmental DNA via three methods: use of degenerate "universal" PCR primers and sequencing of products, high throughput pyrosequencing of tagged rRNA (Andersson et al. 2008), or random sequencing of community DNA with subsequent analysis only of reads containing rRNA (Rusch et al. 2007, Biers et al. 2009). The first method has been demonstrated to be biased due to systematic errors caused by the primers (Manichanh et al. 2008), the second is highly promising for deep assessment of prokaryotic diversity (Huber et al. 2007), and the third uses only a tiny proportion of available reads (considerably less than 1 %). All the approaches have the limitation that they can only provide information about the number of species in a community. Furthermore, the incomplete rRNA lengths generated do not allow assignment to species level, but to higher taxonomic units only (Hanage et al. 2006). Assignment of randomly sequenced reads to a reference genome or gene can lead to further information, such as (given sufficient reads and hits) estimation of the functional capacity of the microbial habitat under study.

Here we provide a brief summary of the potential of short reads in metagenomics and demonstrate their utility in reanalysing existing metagenomes.

5.1.1 Short reads in metagenomics

As next generation sequencers started to produce millions of very short reads in parallel a novel potential application became apparent. This entails the random sequencing of DNA fragments from entire microbial communities, and then their alignment or attribution to potentially present species. Modern short read alignment programs are utilised to avoid the prohibitive running times of algorithms such as Blast with millions of reads. An entire 36 bp short read covers little enough of the possible sequence space (i.e. combinatorial variability) to ensure specificity. This fact has been demonstrated by multiple groups in resequencing of the human genome (Wang et al. 2008, Pushkarev et al. 2009), which at 3000 Mb is of similar size to all currently sequenced bacteria. We anticipate short reads will find favour in deep sequencing and monitoring of metagenomic communities since they a) have the lowest cost per base pair b) are less likely to have large numbers of SNPs and indels due to their short length (Li and Durbin 2009) and c) produce the millions of reads necessary for reliable detection of low abundance species (Sogin et al. 2006, Huber et al. 2007, Andersson et al. 2008). Their biggest advantage, the millions of reads generated, translates into a large sampling effort, meaning that even low abundance species should contribute sufficient reads to be readily distinguished from background noise.

5.2 About the manuscript

This manuscript diverges slightly from the mid-length oligonucleotide approach of this thesis by using longer oligonucleotides, actually short reads, of the order of 30 bp. This figure is arrived at from a 36 bp read with 4 bp trimmed from the typically low quality 3' end, and allowing up to 3 mismatches. The approach detailed here, Metatie, aligns reads to a reference of all sequenced bacteria using the short read aligner Bowtie (Langmead et al. 2009) and various helper applications in a computational pipeline. The main message of the manuscript is that short reads possess sufficient taxonomic signal to be aligned to all currently sequenced bacteria without excessive scattering. A high level of taxonomic accuracy is demonstrated with two artificial genomes, where metagenome composition is known and can be strictly compared to results, and also in two single genome resequencing experiments. Attribution of reads is as high as 99 % when the whole genome has already been completely sequenced, as is the case with *E. coli* K-12. Even metagenomes dominated by sequenced bacteria, such as one Sargasso sea sample site, still return hits with 50 % of reads. These results firmly demonstrate that short reads contain sufficient taxonomic signal to be

of use in the metagenomic arena. Reads are also recovered from artificial metagenomes at strain level in the stoichiometric proportions in which they were inserted. This fact demonstrates the quantitative and highly specific nature of the approach. We advocate using a detection threshold of 100 reads, since 100 aligned short reads are sufficient to estimate sequence length and abundance of a taxon, operon or genomic island in a metagenomic dataset (detailed and discussed fully in the manuscript supplementary materials). Modern short read alignment programs are also very rapid, enabling alignment of millions of reads (less than 15 minutes for 6 million reads; Langmead et al. 2009, Li et al. 2009). Critically, we make use of all developing collections of unfinished bacterial sequences such as GEBA (<http://www.jgi.doe.gov/programs/GEBA/why.html>), Metahit (<http://www.sanger.ac.uk/pathogens/metahit/>), the Moore Microbial Sequencing Programme (<http://www.moore.org/marine-micro.aspx>), and the Human Microbiome Project (Turnbaugh et al. 2007) in addition to the entire NCBI RefSeq project to define the reference. Use of these genomes leads to new insights when analysing new genome collections. For example, we report the occurrence of various newly cultured and sequenced anaerobes in the Japanese human gut metagenome (Kurokawa et al. 2007). For example, the anaerobe *Gordonibacter pamelaee* was detected, though not mentioned in the original study, which had used BlastP analysis of genomes available in 2006. Furthermore, two genera which Metatie found to be among the most dominant, *Alistipes* and *Akkermansia*, were not detected by Kurokawa and coworkers (2007), as they were yet to be sequenced. We also demonstrate how rRNA and randomly sequenced reads attributed by Metatie to various taxonomic groups from the Sargasso sea metagenome correlate. Metatie attributes of the order of 10-30 % of reads from these metagenomes to known species, providing an additional advantage over rRNA, where less than 0.2 % of reads may be used for identification (Biers et al. 2009). Furthermore, the short reads attributed by Metatie provide, in contrast to rRNA, information on the genomic regions and actual genes present in the rich Sargasso dataset. A future project involves adapting Metatie to work on a reference of genes, as opposed to genomes, so as to enable assays of the metabolic capacity of a metagenome.

This manuscript had been submitted to Genome Research at the time of writing. I designed the project, wrote the manuscript and contributed programming and analysis. Jens Klockgether helped with validation and analyses, while Burkhard Tümmler provided mathematical and supervisory assistance. Jens Neugebauer and Ben Langmead programmed various facets.

5.3 References

- Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P. & Engstrand, L. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* **3**(7) e2836.
- Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75
- Biers, E. J., Sun, S. & Howard, E. C. (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75**(7) 2221-2229.
- Brady, A. & Salzberg, S. L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*. Epub August 2.
- Hanage, W. P., Fraser, C. & Spratt, B. G. (2006) Sequences, sequence clusters and bacterial species. *Philos Trans R Soc Lond B Biol Sci* **361**(1475) 1917-1927.
- Huber, J. A., Welch, D. B. M., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A. & Sogin, M. L. (2007) Microbial population structures in the deep marine biosphere. *Science* **318**(5847) 97-100.
- Hugenholtz, P. & Tyson, G. W. (2008) Microbiology: metagenomics. *Nature* **455**(7212) 481-483.
- Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**(4) 169-181.
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3) R25.
- Li, H. & Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*. **25**(14) 1754-60.
- Manichanh, C., Chapple, C. E., Frangeul, L., Gloux, K., Guigo, R. & Dore, J. (2008) A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res* **36**(16) 5180-5188.
- McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**(1) 63-72.
- Pushkarev, D., Neff, N. F. & Quake, S. R. (2009) Single-molecule sequencing of an individual

human genome. *Nat Biotechnol* Epub August 10.

Rappé, M. S. & Giovannoni, S. J. (2003) The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369-394.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**(3) e77.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* **103**(32) 12115-12120.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. (2007) The human microbiome project. *Nature* **449**(7164) 804-810.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., et al. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65.

Metatie - a fast, accurate and adaptable metagenomic classifier for ultra-short reads

Colin F. Davenport¹, Ben Langmead², Jens Neugebauer¹, Jens Klockgether¹ and Burkhard Tümmeler¹

¹ Clinical Research Group, OE6711, Hanover Medical School, Carl-Neuberg-Strasse 1, D-30625 Hanover, Germany.

² Department of Biostatistics, John Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA.

Corresponding author: Colin Davenport, email: davenport.colin@mh-hannover.de

Keywords: metagenome, classifier, short reads, Metatie, Illumina

Abstract

Studies of microbes from various environments have produced large quantities of data leading to great insights into the uncultured microbial world. We apply the speed and sensitivity of modern short read aligners to metagenomic data by using concatenated bacterial genomes as a reference sequence. In addition to completely sequenced bacteria, we include draft sequences from the Human Microbiome and Moore microbial sequencing projects, among others. Our approach, Metatie, is computationally undemanding and enables identification of strains, species and genomic islands from metagenome sequence datasets generated by inexpensive high-throughput short read sequencing technologies. To our knowledge, this method is the only metagenomic approach rapid enough to be feasible for analysing millions of short reads on a single computer. Artificial metagenomes portray the low rate of false positives of our approach at a taxonomic resolution which exceeds existing methods. Metatie is capable of quickly reanalysing existing datasets such as the Sargasso sea metagenome. Short reads contain sufficient phylogenetic signal to obtain similar results to original analyses with longer reads in a fraction of the time. Reevaluation of a Japanese human gut metagenome provides novel insights into gut microbial distributions by detecting newly sequenced anaerobes. There is no trade-off in accuracy for the speed increase shown by Metatie, only the proportion of attributions decreases in contrast to BlastP when using short instead of long reads. In future increasingly diverse microbial communities will become more amenable to exploration facilitated by short read sequencing technologies. Metatie is simple, adaptable and open source.

[Supplemental material is available online at <http://www.genome.org>. A Metatie test server is available at genomics1.mh-hannover.de/metatie/. Programs and source code are available from the authors.]

Introduction

Metagenomics, the analysis of microbial communities from mixed environmental samples (Handelsman 2004), is continually growing in importance. Over 190 metagenomic projects are, in September 2009, currently listed on the Genomes Online database (Bernal et al. 2001). Gaining insights into microbial population and community ecology is now within the reach of many research groups, due to the continually decreasing costs of sequencing. However, many metagenomic studies to date have focussed on the established Sanger sequencing technology, which produces long reads at a high cost per base pair (Rusch et al. 2007, Hugenholtz and Tyson 2008). This phenomenon has dictated that many metagenomes have only been sampled by circa 100,000 reads, which adversely affects the detection of rare organisms (Sogin et al. 2006). As a consequence the Minnesota soil metagenome (Tringe et al. 2005) has been termed uncharacterisable, due to the huge diversity of organisms present (McHardy et al. 2007). Even the current method of choice, Roche 454 pyrosequencing, only typically generates less than a million reads and a total of 0.5 gigabases of sequence data (<http://tinyurl.com/l3jq2o>). Experience from the field of ecology indicates that low abundance organisms, rather than the dominant taxa, may be key providers of ecosystem services (Christianou et al. 2005). If we are to truly understand complex microbial communities, sufficient sampling effort to detect the majority of low abundance species must be routinely applied (Huber et al. 2007). Illumina or similar sequencers allow deeper sampling of microbial samples, with circa 20 million reads per flowcell channel, and 5-25 gigabases in total (<http://tinyurl.com/lup3hv>). The utility of these short reads has already been demonstrated in vertebrate resequencing projects (Bentley et al. 2008), and in a simulated metagenomic context (Warren et al. 2009). Deep sequencing of diverse microbial and picoeukaryotic (Piganeau et al. 2008) communities with these technologies will lead to a more detailed and probably less biased (genes toxic to cloning are underrepresented in clone library based genomic and metagenomic datasets; Sorek et al. 2007) level of understanding.

Ultra short reads from the Illumina sequencing platform have now been available for several years, yet apart from one study (Sorber et al. 2008) to our knowledge have not been applied in metagenomics. This fact stems from the restricted capability of short reads for discovery of novel contiguous DNA such as ORFs, but also from the lack of effective computational approaches for analysing metagenomic short reads, as existing approaches are not fast enough for alignment of millions of reads shorter than 100bp. Early metagenome analysis tools mainly focussed on longer Sanger reads assembled into contigs longer than 1000bp, which increases the accuracy of taxonomic assignment with some methods (McHardy et al. 2007, Huson et al. 2007). However this step concomitantly reduces the number of observations of the microbial community (Biers et al. 2009), as many reads are grouped into a single contig. This is a suboptimal approach if diversity of a microbial community is to be accurately surveyed, as has been recognised by avoidance of assembly in most recent approaches (Dinsdale et al. 2008). Although considerable interest surrounds the development of tools for taxonomic assignment of short reads, few systems are to our knowledge currently available and effective for short reads. One exception is iSSAKE (Warren et al. 2009), a computationally expensive cluster based approach which reports considerable accuracy with simulated short reads but has yet to be tested on real data. Another classifier, PhyMM, can assign 100bp reads at genus level with over 30% accuracy (Brady and Salzberg 2009), the highest yet reported, but is still only of similar speed to Blast.

Classical metagenomics aims to study microbial communities *in situ*. This approach has achieved, and will continue to bring considerable insights into the composition and function of microbial communities. Particularly gene-based surveys have led to advances in understanding of community function. However, nucleotide sequence based analyses encounter problems with the diversity of environmental microorganisms. Most bacterial phyla are underrepresented in the dataset of currently sequenced microorganisms, which is biased towards human pathogens (Huson et al. 2007). While efforts to improve the state of reference genomes are progressing rapidly, another aspect of metagenomics has been less appreciated: that of the importance of precisely monitoring known bacterial species and strains.

Existing metagenomic tools attribute species and strains to non-specific taxonomic groups

at higher taxonomic levels. For example, few sequence reads can be precisely attributed by current programs, even at genus level (Krause et al. 2008). Attributions to higher taxonomic levels, while indicative of the presence of unsequenced phyla in the metagenomic sample, are not an optimal solution. An additional cause for concern is that taxonomy becomes increasingly arbitrary with distance from the species unit (Konstantinidis and Tiedje 2005). Even within a genus, bacteria with distinct lifestyles, pathogenicity and metabolic capacity are grouped together. At least within a monitoring context, greater precision of read attribution to species or strain level is thus a necessity. As monitoring approaches and reference sequences continue to expand, algorithms which are able to rapidly detect species in large metagenomic datasets will become essential. The Blast algorithm, while very effective and the mainstay of some popular metagenomic systems (Huson et al. 2007), will not be able to keep pace.

We here present Metatie, a simple, fast and accurate system for assignment of short reads from prokaryotic metagenomes which can be run on a single computer or via our webserver. This system allows rapid analysis of the vast datasets generated by next generation sequencers and thus facilitates investigation of complex microbial communities at a greater level of detail. We demonstrate the utility of the approach in assessment of phylogenetic origin of millions of reads from two real Illumina genomic datasets, then further verify the utility of 36bp reads using two artificial datasets. Metatie achieves a high proportion of read assignments, with few false assignments to species or strains not in the dataset. Finally we assess real datasets with this algorithm and show its speed and accuracy in reanalysis of serum, gut and ocean habitats. We show that the increase in speed does not come at the expense of accuracy, just a decrease in the proportion of attributions relative to BlastP when using short instead of long reads.

Results

Metatie workflow development

1211 publicly available prokaryotic genome sequences from five sources were concatenated and used as a reference sequence for the short read aligner Bowtie (Langmead et al. 2009). The

Metatite workflow consists of preprocessing and summarising modules which allow rapid assessment of the large alignments generated by Bowtie. The workflow requires less than 20 minutes on an Illumina dataset (6.5 million reads) using 4 processors of a single server.

It is possible to run Metatite with two different tested sets of alignment parameters, in strain specific or species specific modes. Strain specific mode reduces the number of false positives by just reporting reads with only one alignment to any reference genome. In other words, only strain specific regions are detected. If many strains of one species have been sequenced, for example as in *E. coli* or *Prochlorococcus marinus*, then the degree of genomic overlap between strains is likely to be high. Strain unique DNA is hence rare and most reads will not be assigned to a taxon. In contrast, species specific mode reports the best hit of reads with more than one alignment to the reference. This mode allows the accurate attribution of many more reads but does lead to slightly more scatter and false positives, especially incorrect strain attribution (Table 1, Supplementary Figure S2). Both can be used in tandem to reveal distinct information about a dataset.

Accuracy

We tested, in order of increasing complexity, short read genomic resequencing projects, multiple simulated controls, and published metagenomic datasets to assess the sensitivity and specificity of Metatite (Table 1).

Pseudomonas aeruginosa strain CHA

P. aeruginosa strain CHA, a highly virulent clinical isolate from a CF-patient, shows high cytotoxic activity to human neutrophils (Dacheux et al. 2000). We sequenced the CHA genome by Genome Analyzer sequencing-by-synthesis technology, generating more than 6.5 million reads. The highest numbers of hits were to *P. aeruginosa* strains and were assigned 5.7 million (87%) reads (Figure 1). Significantly fewer were attributed to related *Pseudomonas* species, followed by unrelated strains harbouring genomic islands also common to *P. aeruginosa*. Like other *P. aeruginosa* strains, CHA harbours a conserved species-specific core genome and a highly variable accessory genome comprising genomic islands, phage-like sequences and other mobile elements (Mathee et

al. 2008). DNA in the accessory genome is widely acknowledged to have been taken up from other bacterial strains and species (Juhas et al. 2009). For example, 382 reads from the CHA genome were most similar to an 88kb genomic island found in both *P. aeruginosa* strain PACS2 (RGP27, see Mathee et al. 2008) and the beta-proteobacterium *Herminiimonas arsenicoxydans*.

Because one of the main challenges of metagenomic analysis is the fact that species are frequently encountered which have never before been sequenced, we examined Metatite performance after excluding organisms from the reference sequence related to those from which query reads were sampled. Firstly, the four *P. aeruginosa* genomes were removed from the reference. The absolute number of reads attributed decreased from 5.8 million (89%) to 1.2 million (18%), showing Metatite generally avoids making assignments when the species is not known. While more reads are then attributed to related species such as *Pseudomonas mendocina*, this effect is limited in scale to orthologues common to both species (Figure 1). When all *Pseudomonas* species are removed from the reference, the number of CHA reads attributed decreases markedly again to 670,000 (10%). Interestingly, one non-Pseudomonad, *Azotobacter vinelandii* DJ, becomes the top hit with 192,000 reads attributed to it alone. This recently sequenced species was previously known to be very similar to the Pseudomonads (Setubal et al. 2009), and based on 16S ribosomal RNA, whole proteome comparisons and the evidence here, will likely be reassigned to the genus *Pseudomonas* (D.W. Ussery, pers. comm.). In summary, Illumina reads for *P. aeruginosa* were either assigned to strains from the species itself, or when *P. aeruginosa* was removed from the reference, far fewer were assigned, mainly to orthologous regions of related species.

Given the ubiquitous nature of DNA in the environment, detection of contaminant reads from non-target organisms remains an important quality control. We were surprised to detect 182 reads from *Propionibacterium acnes* in our *P. aeruginosa* CHA sample. This bacterium is closely associated with sebaceous human skin (Grice et al. 2009), and was not detected in other *P. aeruginosa* sequencing samples carried out at the same time. We suspected low level human contamination in this case. Aligning reads from the *P. aeruginosa* CHA project to a human genome reference sequence revealed 5406 hits, from all chromosomes. In contrast, only a total of 100

reads from a simultaneously conducted *P. aeruginosa* PAO1 resequencing project mapped to two human chromosomes, and *P. acnes* was not detected (data not shown). This example thus contains two lines of evidence (*P. acnes* and *H. sapiens*) for human contamination and a negative control from a presumably uncontaminated *P. aeruginosa* PAO1 sample.

Escherichia coli K-12

An *E. coli* K-12 strain was resequenced with 10.4 million short reads by Illumina. In species specific mode Metatier assigned 10.2 million (98%) reads to 22 *E. coli* strains and 7 *Shigella* species, with *E. coli* K-12 and ATCC8739 having the most reads assigned (4.3 million, 41%). *Shigella flexneri* and *E. coli* K-12 are extremely similar, sharing a common genomic backbone (Jin et al. 2002). It is also known that *E. coli* K-12 shares a number of orthologues with *Salmonella*, and *Citrobacter koseri* contains multiple loci of about 30kb in size which share 90% identity with *E. coli* K-12. These facts explain why 12,904 reads were mapped to *C. koseri* and up to 11,094 to various *Salmonella* strains. Metatier's strain level approach, where reads with only one good hit to the entire reference are reported, only attributes 639 (0.1%) reads. This occurs because *E. coli* strain K-12 is present in the reference sequence multiple times, so very few strain specific regions are present.

For comparison, in an analysis of 2,000 100bp 454 reads, the metagenomic classifier Megan correctly assigned 1,052 (53%) to the Enterobacteriaceae family. 542 reads (27%) were not assigned and the remaining 510 reads were assigned at less informative taxonomic levels. Only 128 reads (6%) could be attributed correctly to *E. coli*, and just 31 (2%) could be assigned to *E. coli* K-12. By way of comparison, Metatier could assign 4,316,132 (41%) shorter 36bp reads to *E. coli* K-12 (Supplementary Figure S4). This effect is not due to presence or absence of strains in the respective reference sequences, but on the specificity of the classification approach. It is probable that Megan's lowest common ancestor algorithm (see Huson et al. 2007) was disadvantaged by the taxonomic grouping of the very closely related *Shigella* and *E. coli* strains in two distinct genera. Further examples which exemplify the strength of Metatier in contrast to other systems in providing effective species or strain level assignment are detailed in the discussion.

Artificial metagenomic datasets

We constructed a simple dataset mimicking an ocean environment to test Metatie. Errors were introduced based on a pyrosequencing model to mimic those made by sequencing (Richter et al. 2008). All of the ten strains included were found by Metatie in species level analysis mode in the same stoichiometric proportions as entered (Table 1, Supplementary Figure S1). Next, a published dataset comprising 113 strains with introduced simulated errors (Mavromatis et al. 2007) was used to validate Metatie performance. A detection threshold of 100 reads was utilised because 100 aligned short reads are sufficient to estimate sequence length and abundance of a taxon, operon or genomic island in a metagenomic dataset (see equations 3-5 in the Supplementary Material). 101 of 113 strains (89.4%) were detected in species level mode (Table 1, Supplementary Figure S2). Seven strains were not detected (false negatives). Of these, four were not in the Metatie reference dataset, while three were below the detection threshold. No strains included in the metagenome were missed completely by Metatie. For the five remaining taxa, strains for four species were wrongly predicted and the species for one genus (4%) was falsely predicted to be present. The accuracy of these results are reinforced by the fact that all three chromosomes of each *Burkholderia* strain were separately recovered.

Seven strains of the dominant alphaproteobacterial species, *Rhodopseudomonas palustris*, were then excluded from the Metatie reference and the analysis was rerun. Our goal was again to check whether Metatie would incorrectly attribute *R. palustris* reads to other species as false positives. Exclusion of *R. palustris* caused the total number of reads attributed to drop by 24,417 (84% of the reads originally attributed to this species). A small increase in reads attributed to the related alphaproteobacteria *Bradyrhizobium* and *Nitrobacter* was noted, probably due to orthologous genes (Supplementary Figure S2). In other words, DNA from unknown species in a metagenome remains mostly unassigned, with a small proportion attributed to related taxa.

Real metagenomic datasets

Having verified the specificity and thus suitability of 36bp reads for use in a metagenomic scenario, we then applied Metatie to three metagenomic datasets of increasing complexity.

Illumina human serum metagenome dataset

Metatier was used to assess the taxonomic content of Illumina 36bp reads from DNA in serum from a patient with hepatitis B virus related liver failure (Sorber et al. 2008). A considerable proportion of reads (14.8%) were naturally found to be human rather than viral or bacterial in nature (Table 1). Of the 3289 viruses in the viral reference sequence, only the hepatitis B virus with 635 reads was reliably detected (Supplementary Table S1). This result illustrates the specificity of 36bp short reads once again. The original authors recruited 328 reads using only this virus as a reference, with the difference presumably due to alignment settings. A few reads were fascinatingly predicted to be from bacteria, although serum is highly bactericidal. The bacterium detected, *Propionibacterium acnes*, inhabits human skin (Grice et al. 2009) and presumably either entered the bloodstream with the puncture wound and was immediately lysed, or represents very low level contamination through handling as was observed with *P. aeruginosa* CHA.

Sargasso sea stations 13 and 11 sample GS000a

Sanger reads from a Sargasso sea sampling site (Venter et al. 2004), a more complex environmental metagenome, were investigated next. Metatier could assign 289,442 reads (44.9%) of this dataset to bacteria (Table 1). Various *Burkholderia* and *Shewanella* strains dominated the sample (Figure 2, Supplementary Table S2). More typical oceanic strains such as *Aeromonas*, *Prochlorococcus*, *Synechococcus* and *Candidatus Pelagibacter* were also detected in considerable numbers. Results were compared to those from ribosomal RNA (Figure 3). Biers and coworkers (2009), in an extremely thorough investigation of Global Ocean Sampling data using 16S rRNA, could assign just 0.12% of these reads to rRNA. 90% of reads could be classified to approximately family level. Comparison of rRNA and Metatier observations is difficult due to the differential number of rRNA operons in taxa (Rusch et al. 2007). However, normalised Metatier and rRNA assignments to the main taxa are roughly comparable (Figure 3). The observed variability is similar to that from two different groups which both assessed the complete Global Ocean Sampling metagenome using rRNA (Supplementary Figure S3). The SAR11 clade is currently better covered

by rRNA sequences than by sequenced genomes, hence rRNA attributions exceed Metatier attributions for this dataset. Five representative rRNA operons were used by Biers et al. (2009), yet only one genus, *Candidatus Pelagibacter*, to our knowledge has sequenced members.

Prochlorococcus sp., and subsequently all Cyanobacteria, were then removed from the Metatier reference as detailed previously. The total proportion of reads attributed decreased as expected, from 44.9% to 43.4% and then to 42.4% (Figure 2). Exclusion of either *Prochlorococcus* or all Cyanobacteria had no effect on the number of reads assigned to other taxa (such as *Aeromonas*, *Burkholderia*, "*Candidatus Pelagibacter* sp.", or *Pseudomonas*).

Reanalysis of a Human Gut metagenome

Lastly, we pooled reads from 13 patients of a thorough human gut metagenome project (Kurokawa et al. 2007). It was not our intention to reanalyse the biological meaning of this dataset, but to test Metatier with a large quantity of real clinical data. Kurokawa and coworkers inferred taxonomy using a BlastP based approach with predicted protein coding genes (alignment over 50% of query length with at least 90% identity). Metatier found all genera inferred by Kurokawa et al. (2007) with the exception of *Rauvolfia*, which is not present in the Metatier reference (Figure 4, Supplementary Table S3). Additionally, Metatier found a whole range of numerically significant species, such as *Akkermansia*, which Kurokawa et al. did not report. Some of these are novel anaerobes which were presumably not sequenced in 2007. Metatier also predicted that the archaeon *Methanobrevibacter smithii* was present in the dataset, contrary to the findings of Kurokawa and coworkers who found virtually no archaea in Japanese subjects. However Metatier's prediction is in line with previous work by other researchers on American subjects (Gill et al. 2006). It does seem intuitive that methanogenic archaea are present in gut metagenomes (Zhang et al. 2009).

Our approach, using 36bp read fragments, is not able to assign the same proportion as BlastP using full length 800bp Sanger reads (Figure 4). Metatier typically assigns about a quarter as many reads as the BlastP approach for this dataset with the alignment settings applied here to comparable or more specific taxonomic levels. However, the increased numbers and rapidly

increasing lengths of reads produced by short read sequencers make this issue largely irrelevant.

Discussion

Resequencing analyses from large scale projects such as the 1000 genomes project (Kaiser 2008) have already demonstrated the potential and reliability of aligning short, high-quality sequences to reference genomes in the order of 3 gigabases (Bentley et al. 2008), which is of similar size to all public completed prokaryotic genomes at present. Using concatenated microbial sequences as a reference sequence for the short read aligner Bowtie has allowed us to extend the scope of algorithms intended for use in resequencing vertebrate genomes to the metagenomic field. Short reads are sufficient for fast, definite assignments without the ambiguity seen with long read alignments and the introduction of gaps. Furthermore, unpaired 36bp reads are the cheapest form of sequence information currently available (Hugenholtz and Tyson 2008), which should facilitate their usage in metagenomic monitoring projects. As read lengths change and paired end reads are used, the modular nature of Metatie means modifying Bowtie alignment parameters (see Supplementary Material for guidelines) or substituting an alternative alignment tool into the Metatie pipeline is a simple task.

The sensitivity of Metatie improves greatly on other currently available classifiers analysing short reads, allowing attribution of large numbers of reads at species or strain taxonomic level. Firstly, the classic 16S ribosomal RNA gene is ineffective with randomly sequenced short reads due to its high degree of conservation. Even with long reads, 0.2% or less typically consist of ribosomal RNA (Biers et al. 2009). In practice, whole phyla are predicted to be present based on two or three reads which hit rRNA, and two groups assigning rRNA can come to very different findings (Supplementary Figure S3; Biers et al. 2009). Furthermore, the order of magnitude copy number variations of rRNA operons (Klappenbach et al. 2001) make this approach effective for

qualitative, but unsuitable for quantitative assessment of microbial communities (Rusch et al. 2007, Manichanh et al. 2008). The classifier Phylopythia only functions on contigs (assembled reads) with sequence lengths in excess of 1000bp and yet still fails to assign 98.3% of contigs from the Sargasso sea sample at genus level (McHardy et al. 2007). Another algorithm, CARMA (Krause et al. 2008), also shows limitations at a specific phylogenetic level, assigning just 6% of all input reads accurately at genus level. Phymm (Brady and Salzberg 2009) is a composition based classifier which reports the highest accuracy for short (100bp) reads at genus level (32.8%). However, these results are for artificial reads without any introduced errors, and 67% of all reads were misclassified, as all reads were assigned. Moreover, Phymm was further tested on the very low complexity acid mine metagenome, which harbours less than 10 species (Tyson et al. 2004) and is thus too simple to demonstrate Phymm's accuracy on a real environmental sample. Finally, the well known BlastP is an accurate algorithm which could assign a large proportion of 800bp reads to various taxa (Figure 4; Kurokawa et al. 2007). While Metatier could not assign as many 36bp reads as BlastP 800bp reads, this limitation is more than compensated for because short read sequencers generate several orders of magnitude more reads than the Sanger technique.

It is difficult to comprehensively evaluate the performance of metagenomic classification tools without the use of carefully controlled datasets. The simulated low complexity dataset is fairly complicated with a high number of strains (113), and the Sanger simulated reads are not particularly well suited to testing a metagenomic tool intended for interpretation of millions of short reads. Nevertheless, Metatier performed very well, assigning a very high proportion of reads to retrieve correct strains in 89% of cases. The sensitivity of the method even allows identification of multiple chromosomes from strains possessing more than one replicore. Performance on the simple in-house simulated ocean dataset was also good. Reads were recovered from all strains in stoichiometric ratios similar to the number inserted. In species level detection mode, only the included or closely related strains were detected. Strain level assignment mode eliminated false positives. While attribution of reads to the sequenced genome from which they came may seem trivial, current metagenomic programs (Brady and Salzberg 2009) do not achieve this with the same level of specificity as Metatier .

In addition to its increased sensitivity and ability to take advantage of very short reads, Metatier is also far more rapid than existing approaches. The Blast algorithm behind Megan would require over 1000 hours on a compute cluster to align 1.6 million Sargasso Sea reads from the full dataset (Huson et al. 2007). As a consequence Megan's authors only investigated a limited set of reads (20,000), but achieved an assignment for most which agreed well with conceptually distinct analysis by the initial investigators (Venter et al. 2004). However, many of these hits are at a less informative taxonomic level than that reported by Metatier. In other words, there is little observable tradeoff in accuracy for the speed increase shown by Metatier, only the proportion of attributions decreases in contrast to BlastP when using short instead of long reads. Computational time limitations, while already a real problem (Hugenholtz and Tyson 2008), will become prohibitive as metagenomic data becomes more common and routine analysis moves away from large computational resources to that realistic for individual research labs. The computational power needed by BLAST or Phymm is not suitable for the number of reads generated by modern sequencers, especially taking into account the ever increasing amounts of sequence data being generated.

Assessment of unknown DNA remains a key challenge of metagenomics, as the vast majority of bacteria, aside from pathogens, have not yet been sequenced (Rappé and Giovannoni 2003). The above-mentioned programs take the approach of tentatively assigning unknown reads to a higher taxonomic level. However, especially in the prokaryotic world, taxonomy becomes more arbitrary and less informative as one moves away from the species level (Konstantinidis and Tiedje 2005). Furthermore, attributions of reads to a genus are inherently unsatisfying to physicians or biologists, especially given that even single strains within a species may display large variations in function or virulence. Metatier takes an alternative but complementary approach which achieves improvements in specificity, speed and practicality over existing programs. Reads are only attributed when the reference sequence is known, thus unsequenced genomes are explicitly ignored (Figure 1, Figure 2, Supplementary Figure S2). This limitation is universal, as even the most advanced compositional classifiers can only attribute reads to known sequences (Brady and Salzberg 2009), and any other expectation of a classifier is unrealistic. In other words, reads from

unknown (unsequenced) species cannot be reliably distinguished and assigned to taxonomic bins by any algorithm, as the sequence space they occupy is unknown. In any case, availability and taxonomic diversity of reference sequences are rapidly increasing with progress of large scale sequencing initiatives and concomitant decreases in sequencing costs (Turnbaugh et al. 2007). Advances in retrieval of strains and sequencing without culturing, for example by labelling DNA (Kalyuzhnaya et al. 2008), will further accelerate accumulation of sequenced uncultured organisms.

Paralogues and horizontally transferred genetic elements present a challenge to sequence-based metagenomics analyses. We circumvent these problems by reporting, in strain specific mode, only reads which have one good alignment to the entire metagenome reference sequence. As a consequence, if a read aligns to a genomic island with three base pairs or less differences in more than one different species, then no alignment will be reported. On the other hand, the strong phylogenetic signal of short reads allows a quick and comprehensive analysis of DNA sequenced from a single strain using Metatier's species specific mode. For example, the presence of reads attributed to *Herminiimonas arsenicoxydans* found in the *P. aeruginosa* CHA example can be explained by components of a putative genomic island common to both strains (Supplementary Figure S5). In a metagenomic context it is difficult to distinguish the presence of whole genomes (Supplementary Figure S6) or genomic islands from read counts alone. Thus we present several algorithms able to distinguish read distributions in the Supplementary materials.

Short 36 base pair reads, while in theory highly specific, will in practice show a degree of scatter to various taxa, as prokaryotic sequence space contains overlaps such as repeats or duplicated genes (Davenport et al. 2009). Hence we advocate using both a simple threshold and the algorithms listed in Supplementary Materials. According to equations 3-5 the mapping of the short reads onto the concatenated pangenome allows estimation of species abundance and chromosome size in the bacterial metagenome. Moreover, the mapped reads can be differentiated as to whether they represent an already sequenced gene/operon, island or species. In other words, homologs of known genomic segments can be identified in taxa for which no reference genome has yet been sequenced. According to the algorithms mentioned above operons, genomic

islands or chromosomes respectively require, depending on length, about 10-20, 30-50 or 100-200 reads, to be reliably detected.

Given the high number of observations and demonstrated capability for fast and specific alignments, massively parallel short read sequencing is an underutilised yet immensely promising technology for metagenomic application. Furthermore, the algorithms introduced here allow distinction between genomic islands and entire genomes in the mapped data. Metatie's current focus on assessing known diversity can be seen as complementary to other approaches which assess gene and metabolic capacity. However, using a custom concatenated reference sequence of genes (instead of genomes), Metatie could be used to evaluate the presence of genes, and thus community metabolic capacity, using short read data. In light of the results presented here, Metatie represents a capable system for rapid assessment of metagenomes sequenced with short read sequencing technologies (Figure 5).

Methods

The Metatie pipeline

Metatie is built around the short read mapping program Bowtie (Langmead et al. 2009). Bowtie can rapidly map 36bp reads to very long reference sequences. Bowtie was then adapted to a metagenomic scenario: 1211 prokaryotic chromosomes from various sources were concatenated and used to build a metagenomic reference sequence. These sources include May 2009 versions of the NCBI RefSeq collection, the Human Microbiome project (Turnbaugh et al. 2007), the Genomic Encyclopedia for Bacteria and Archaea for sequencing of microbes from poorly represented phylogenetic groups (<http://www.jgi.doe.gov/programs/GEBA/why.html>), the Metahit programme, and the Moore Foundation Marine Microbial Genome Sequencing Project (<http://www.moore.org/marine-micro.aspx>). Short reads can then be mapped onto this reference in a similar fashion to that now routinely used in resequencing of mammalian genomes. Reference sequences have also been built for 3289 viruses and 1725 plasmids from the NCBI RefSeq

collection. Strain genomes could be excluded from reference multi-FastA files using an in-house Perl script.

A set of Perl, Java and R scripts collectively known as Metatie preprocess metagenomic data in FastA or FastQ formats, run the Bowtie aligner, process the output data and graphically summarise output. A Perl script runs the entire workflow. Because short reads offer the cheapest cost per bp and have been underutilised by metagenome researchers to date, 36bp of sequence were filtered from metagenomic projects with longer reads with an in-house script. Short reads were extracted from Sanger reads at position 100 to avoid ambiguous bases present towards the beginning. The Bowtie aligner is subsequently run with default settings, except for a seed size of 26bp (-l 26) with 3 mismatches allowed (-n 3), mismatch quality sum of 100 (-e 100), and 4 bp 3' read trimming (-3 4). In strain specific mode only reads with more than one alignment are suppressed (additionally -m 1). The next step involves creation of a summary text file of all genomes represented by 20 or more read alignments using the Java script MapProcessor, which also adds taxonomy information taken from the headers of each FastA file of the reference dataset at lineage, genus, species and strain level. Results are then plotted using the statistical language R (R Development Core Team, 2008). Finally, a PHP file is produced which reports text summaries and graphical output to the user's browser.

Analysis of existing datasets

The Metatie pipeline was used to analyse the following datasets, which were downloaded from the NCBI Trace archive (<ftp://ftp.ncbi.nih.gov/pub/TraceDB/>) or the CAMERA database (<http://camera.calit2.net/>): *Escherichia coli* K-12 MG1655 36bp reads (NCBI GenBank short read archive accession SRA001125), Sanger reads from the simulated low complexity dataset (Mavromatis et al. 2007), Illumina reads from human serum (Sorber et al. 2008), Sanger reads from Sargasso Sea stations 13 and 11 site GS000a (Venter et al. 2004) and pooled human gut metagenome Sanger reads (Kurokawa et al. 2007). *P. aeruginosa* CHA Illumina 36bp reads were available in this laboratory.

Creation of the simulated ocean dataset

A simple artificial ocean metagenome was created using the program Metasim (Richter et al. 2008) with 100000 reads from ten marine strains: 25076 *Prochlorococcus marinus* MIT 9301 reads, 25032 *Synechococcus elongatus* PCC 7942 reads, 6211 from "*Candidatus* Pelagibacter ubique HTCC1062", 6273 from *Marinobacter aquaeolei* VT8, 6228 from *Marinomonas* sp. MWYL1, 6367 from *Nitrosococcus oceani* ATCC 19707, 6203 from *Nitrosopumilus maritimus* SCM1, 6291 from *Oceanobacillus iheyensis* HTE831, 6110 from *Pelobacter carbinolicus* DSM 2380, and 6209 from *Shewanella baltica* OS155.

Notation for unfinished (non-RefSeq) genomes

The usual RefSeq notation, e.g. NC_000001, was slightly modified to accomodate unfinished genomes which are not yet (and may never be) included in RefSeq.

NC_m00022 source: the Moore microbial sequencing project.

NC_i00019 source: human microbiome project, retrieved from Integrated Microbial Genomes.

NC_j00001 source: GEBA project, retrieved from the Joint Genome Institute.

NC_s00001 source: Metahit project, from the Sanger centre.

P. aeruginosa CHA sequencing

Sequencing of the CHA genome by Genome Analyzer sequencing-by-synthesis technology was performed at GATC-Biotech (Constance, Germany) using the Illumina platform. 6,565,136 36bp reads were generated which fulfilled general quality criteria. During primary analysis with the ELAND (Illumina) alignment tool 5,470,456 reads were aligned to the *P. aeruginosa* PAO1 reference sequence (Stover et al. 2000). Thus, 96.5% of the 6.264 Mb of the PAO1 genome was covered. As the *P. aeruginosa* specific core genome is highly conserved most reads not aligned to the PAO1 reference (16% of all reads) should represent elements of the accessory genome.

Availability

A web server with test datasets is available online (<http://genomics1.mh-hannover.de/metatie/>). To

analyse your own datasets or map onto a custom reference set, please contact the authors to upload your data or obtain the freely available scripts.

Acknowledgements

We are grateful to Frauke Stanke and Lutz Wiehlmann for helpful discussions. GATC (Constance, Germany) sequenced *P. aeruginosa* strain CHA. C.D. is a member of the International Research Training Group 'Pseudomonas Pathogenicity and Biotechnology' sponsored by the Deutsche Forschungsgemeinschaft. B.L. was supported in part by NIH grant R01-GM083873.

Figure legends

Figure 1: Metatite species level attributions of 6.6 million 36bp Illumina reads from a *P. aeruginosa* CHA sequencing project. A full reference, a reference lacking *P. aeruginosa* and a reference lacking all *Pseudomonas* sp. were used to simulate the effect of DNA from unsequenced species in a metagenome. A total of 5.8 million (89%) reads were assigned to the full reference, 5.7 million (87%) to the four related sequenced *P. aeruginosa* strains. Exclusion of *P. aeruginosa* strains from the reference caused a large reduction in assignments, and did not lead to a large increase in scattered false attributions to other species.

Figure 2: Number of 36bp reads from the Sargasso sea stations 13 and 11 (GS000a) dataset assigned by Metatite to marine microbial genera. A full reference, a reference lacking *Prochlorococcus* and another lacking all Cyanobacteria were tested. The main constituents of this dataset are from the genera *Burkholderia* and *Shewanella*. Full numbers of reads assigned to these genera are indicated above the bar, and did not vary when the reference was modified. Manipulation of the reference sequence does not lead to false positive attributions of reads, e.g. to orthologous genes in related genera, when the genus *Prochlorococcus* is removed.

Figure 3: Comparison of Metatite read assignments and ribosomal RNA assignments for the Sargasso sea stations 13 and 11 GS000a dataset. Metatite read proportions were normalised by the total number of reads reliably assigned, while ribosomal RNA data is from Biers et al. (2009). CFB is an abbreviation for the Cytophaga-Flavobacterium-Bacteroides taxonomic group.

Figure 4: Venn diagram of taxonomic assignment of a human gut metagenome project with BlastP (left circle, Kurokawa et al. 2007) versus species level assignment by Metatite (right circle). B indicates the percentage of 800bp Sanger reads assigned by BlastP, while M denotes percentages of 36bp reads attributed by Metatite. One 36bp read fragment was extracted from each of the 1,501,410 Sanger reads, and 12% of reads were assigned by Metatite. Proportions of reads assigned by BlastP were estimated from Kurokawa et al. (2007).

Figure 5: Use cases for the Metatite system in analysing single short read genome sequencing projects or prokaryotes in a metagenomic community.

Tables

Table 1: Summary of Metatite performance using stringent (strain level) and more relaxed parameters (species level accuracy) on genomic sequencing, artificial and real metagenomic datasets projects. A threshold of 20 reads was used, below which hits were not reported.

Dataset	Sequencing technology	Total reads (uncontaminated reads*)	Strain level: Reads assigned (%)	Species level: Reads assigned (%)
<i>Pseudomonas aeruginosa</i> CHA	Illumina 36bp	6,565,136 (6,559,730)	262,646 (4.0%)	5,861,120 (89.4%)
<i>P. aeruginosa</i> CHA vs reference lacking <i>P. aeruginosa</i>	Illumina 36bp	6,565,136 (6,559,730)	406,111 (6.2%)	1,197,370 (18.3%)
<i>P. aeruginosa</i> CHA vs reference lacking all <i>Pseudomonas</i> sp.	Illumina 36bp	6,565,136 (6,559,730)	293,215 (4.5%)	669,480 (10.2%)
<i>Escherichia coli</i> K-12 MG1655	Illumina 36bp	10,408,224 (10,403,764)	639 (0.1%)	10,335,773 (99.3%)
Simulated ocean	Sanger	100,000	18,777 (18.8%)	48,878 (48.9%)
Simulated low complexity	Sanger	97,494	53,562 (54.9%)	82,676 (84.8%)
Simulated low complexity vs reference lacking <i>Rhodopseudomonas</i> sp.	Sanger	97,494	37,478 (38.4%)	58,259 (59.8%)
Human serum Hepatitis B vs viral reference	Illumina 36bp	294,625 (250,979)	270 (0.1%)	678 (0.27%)
Sargasso sea site GS000a	Sanger	644,551 (644,000)	78,364 (12.2%)	289,442 (44.9%)
Japanese human gut	Sanger	1,501,410 (1,500,946)	106,729 (7.1%)	184,888 (12.3%)

*putatively contaminated reads were aligned against the human genome by Metatite and excluded from further analysis. This step was not required for the artificial genomes.

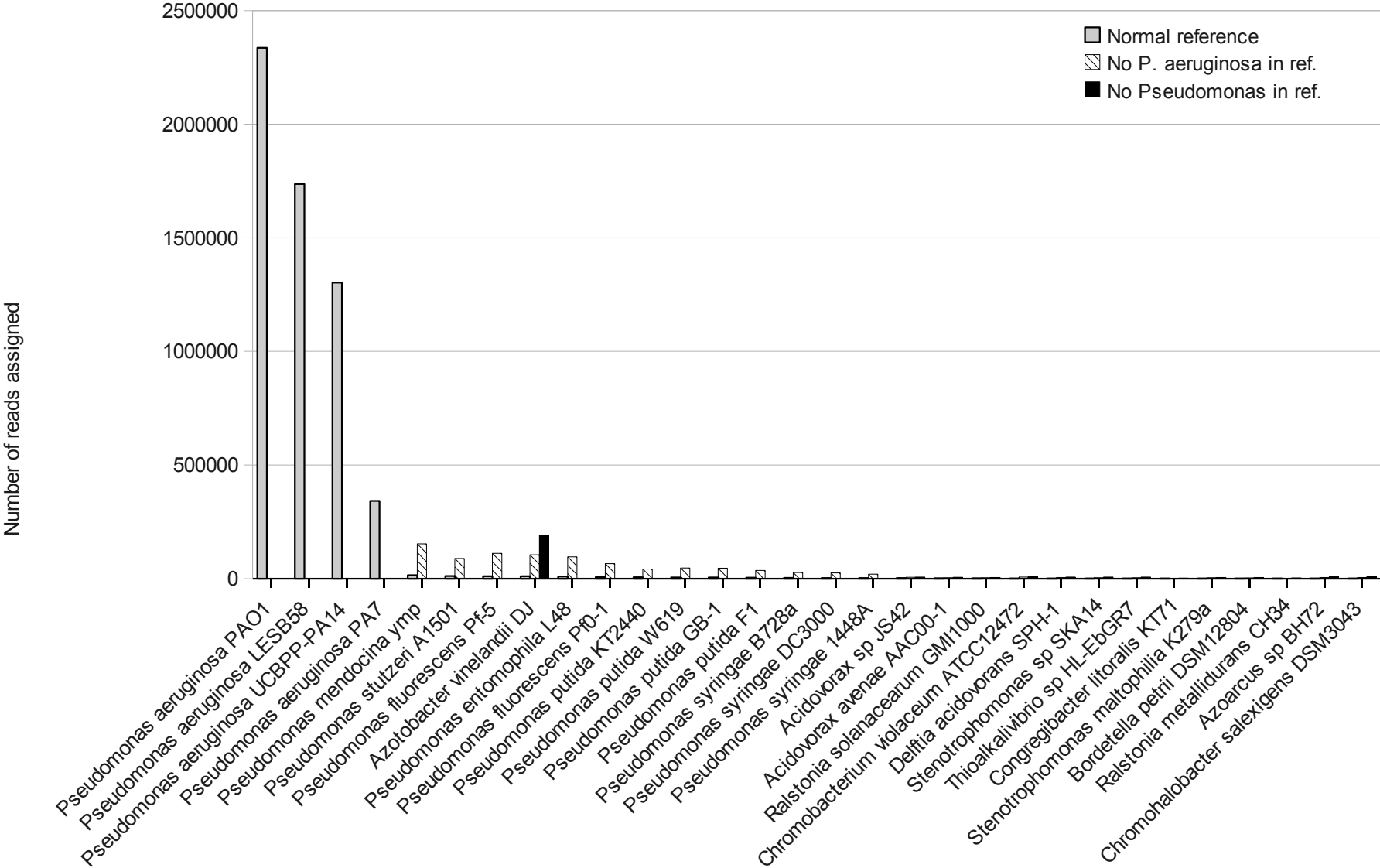
References

- Bentley SD, Chater KF, Cerdeño-Tárraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.
- Bernal A, Ear U, Kyrpides N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29**: 126-127.
- Biers EJ, Sun S, Howard EC. 2009. Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol*. **75**: 2221-2229.
- Brady A, Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*. **6**: 673-676.
- Christianou M, Ebenman B. 2005. Keystone species and vulnerable species in ecological communities: strong or weak interactors? *J Theor Biol*. **235**: 95-103.
- Dacheux D, Toussaint B, Richard M, Brochier G, Croize J, Attree I. 2000. *Pseudomonas aeruginosa* cystic fibrosis isolates induce rapid, type III secretion-dependent, but ExoU-independent, oncosis of macrophages and polymorphonuclear neutrophils. *Infect Immun* **68**: 2916-2924.
- Davenport CF, Wiehlmann L, Reva ON, Tümmler B. 2009. Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ Microbiol*. **11**: 1092-1104.
- Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature* **452**: 629-632.
- Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355-1359.
- Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Program NISCCS, Bouffard GG, Blakesley RW, Murray PR, et al. 2009. Topographical and temporal diversity of the human skin microbiome. *Science* **324**: 1190-1192.
- Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**: 669-685.
- Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. 2007. Microbial population structures in the deep marine biosphere. *Science* **318**: 97-100.
- Hugenholtz P, Tyson GW. 2008. Microbiology: metagenomics. *Nature* **455**, 481-483.
- Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res* **17**: 377-386.
- Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, et al. 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res* **30**: 4432-4441.

- Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* **33**: 376-393.
- Kaiser J. 2008. DNA sequencing, a plan to capture human diversity in 1000 genomes. *Science* **319**: 395.
- Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E, Salamov A, Grigoriev IV, Suci D, Levine SR, et al. 2008. High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol* **26**: 1029-1034.
- Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. 2001. rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res* **29**: 181-184.
- Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258-6264.
- Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, et al. 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**: 169-181.
- Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Manichanh C, Chapple CE, Frangeul L, Gloux K, Guigo R, Dore J. 2008. A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res* **36**: 5180-5188.
- Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, Rokas A, Yandava CN, Engels R, Zeng E, et al. 2008. Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci U S A* **105**: 3100-3105.
- Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4**: 495-500.
- McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos, I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**: 63-72.
- Piganeau G, Desdevises Y, Derelle E, Moreau H. 2008. Picoeukaryotic sequences in the Sargasso Sea metagenome. *Genome Biol* **9**: R5.
- R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rappé MS, Giovannoni SJ. 2003. The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369-394.
- Richter DC, Ott F, Auch AF, Schmid, R, Huson DH. 2008. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE* **3**: e3373.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K. 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**, e77.

- Setubal JC, dos Santos P, Goldman BS, Ertesvåg H, Espin G, Rubio LM, Valla S, Almeida NF, Balasubramanian D, Cromes L et al. 2009. Genome sequence of *Azotobacter vinelandii*, an obligate aerobe specialized to support diverse anaerobic metabolic processes. *J Bacteriol* **191**: 4534-4545.
- Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* **103**: 12115-12120.
- Sorber K, Chiu C, Webster D, Dimon M, Ruby JG, Hekele A, DeRisi JL. 2008. The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE* **3**: e3495.
- Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**: 1449-1452.
- Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrenner P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, et al. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**: 959-964.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC. 2005. Comparative metagenomics of microbial communities. *Science* **308**: 554-557.
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI, et al. 2007. The human microbiome project. *Nature* **449**: 804-810.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.
- Warren RL, Nelson BH, Holt RA. 2009. Profiling model T-cell metagenomes with short reads. *Bioinformatics* **25**: 458-464.
- Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing R, Rittmann BE, et al. 2009. Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci U S A* **106**: 2365-2370.

Figure 1



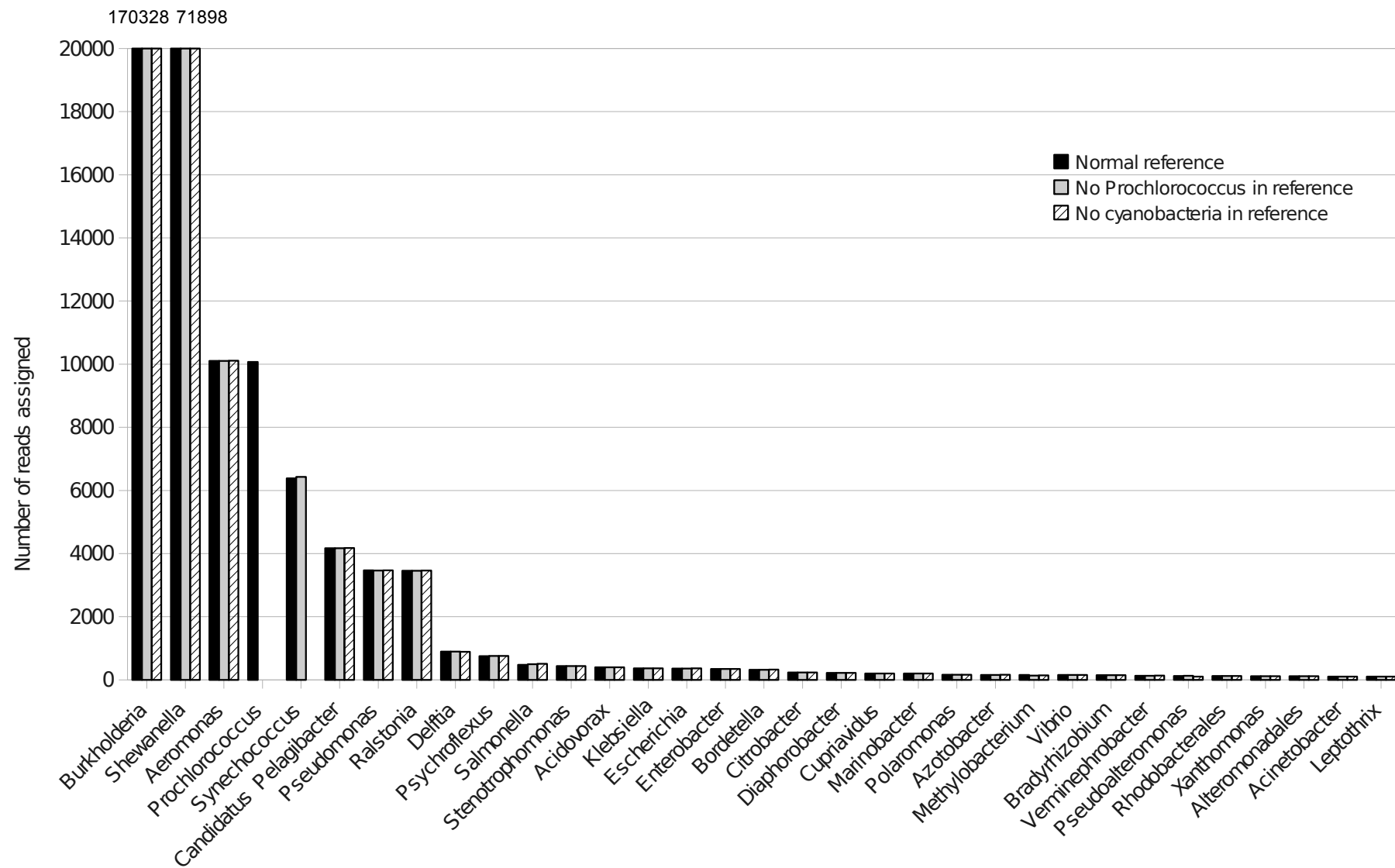


Figure 2

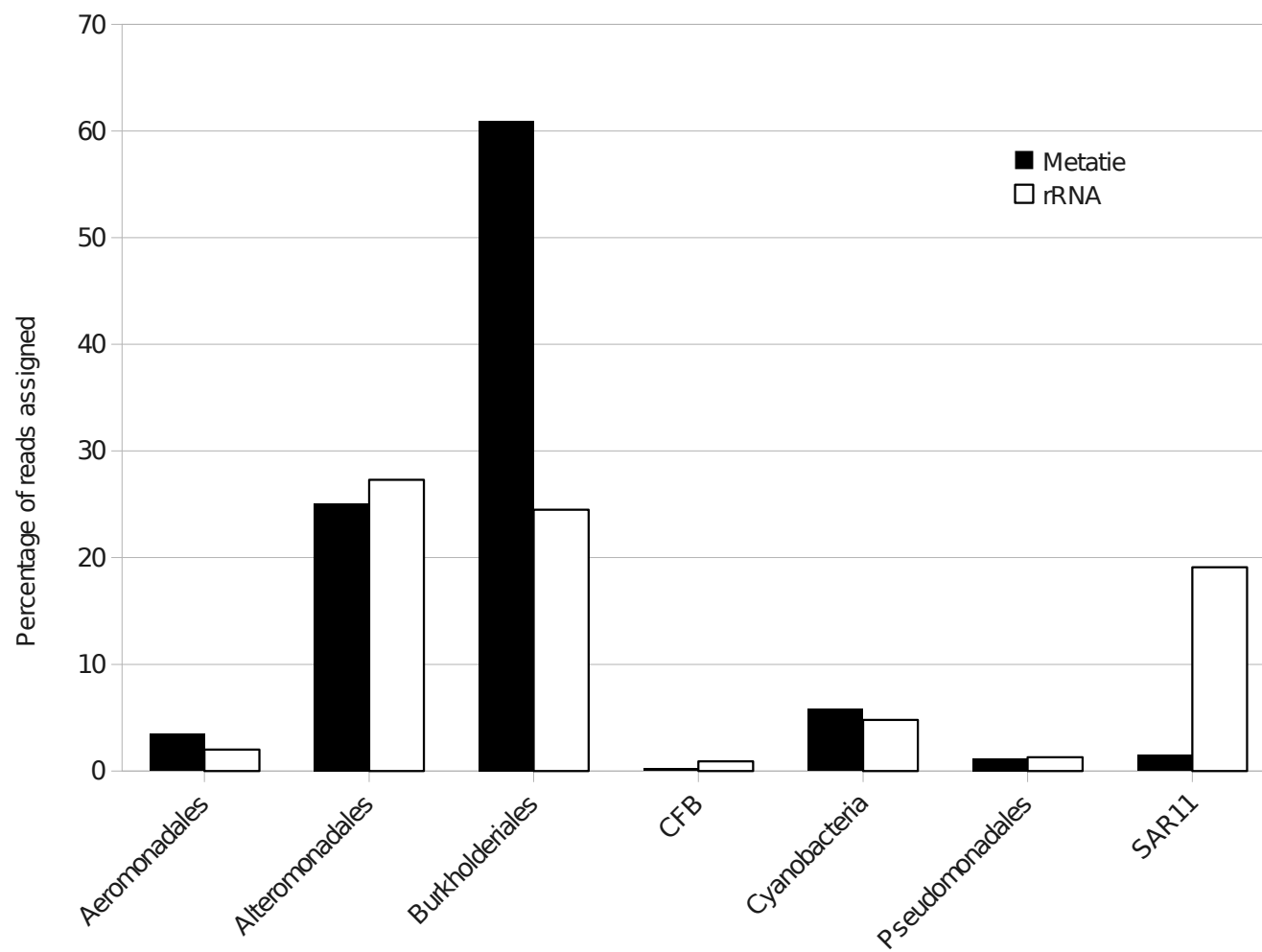


Figure 3

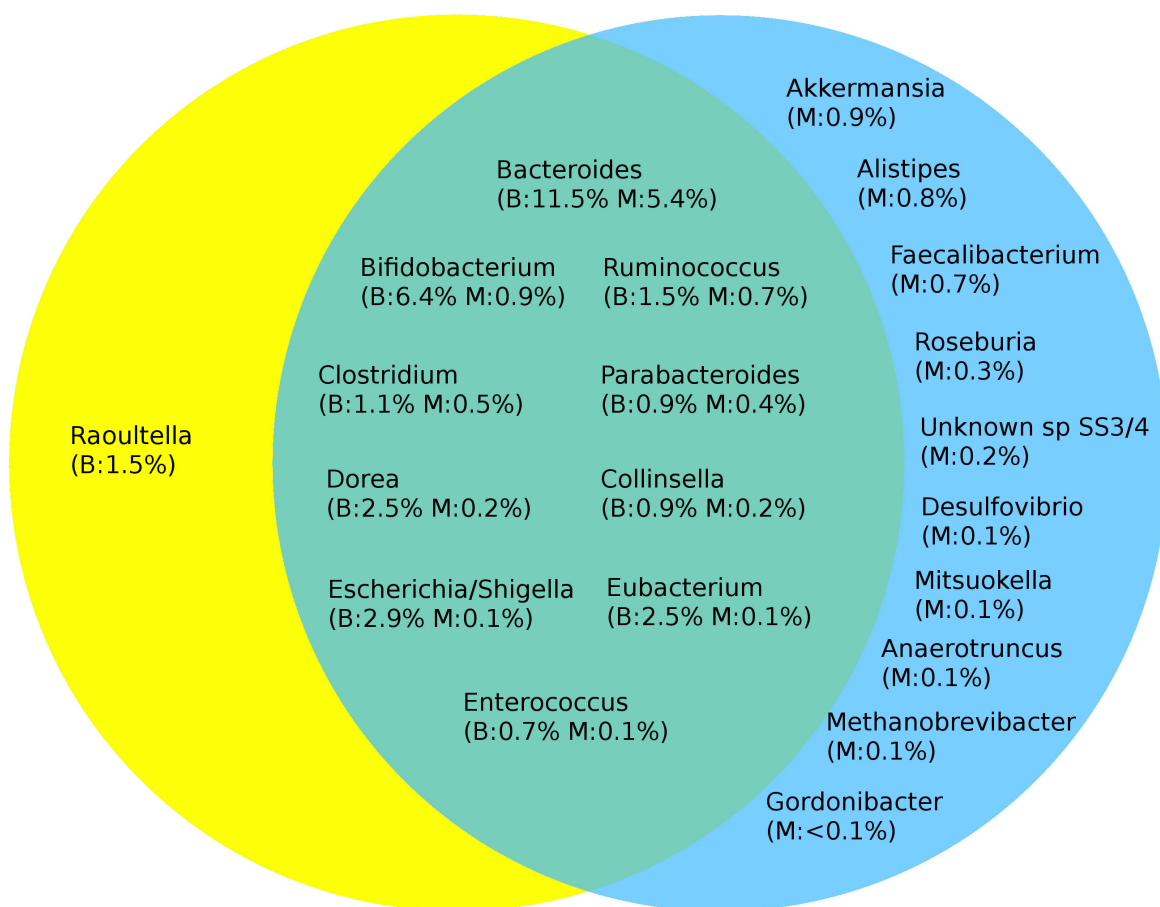
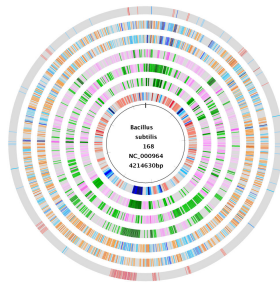
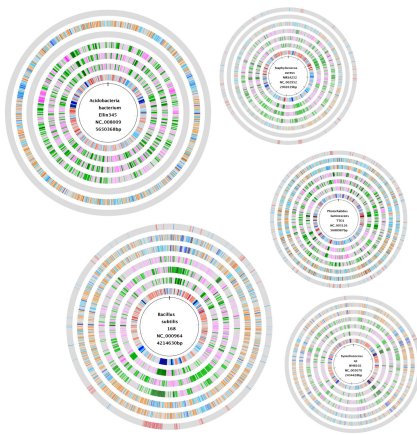


Figure 4



Genomics

- quick analysis of genes and islands present



Metagenomics

- rapid and specific detection of bacteria in sample
- can use 36bp reads

Figure 5

Conclusions and Outlook

1. Oligonucleotide signature analysis

The oligonucleotide sequence analysis presented within this thesis focusses on discoveries of medium sized 8-14 bp oligomers and their usage in three main applications (i) assessment of divergent genomic regions in prokaryotic genomes, (ii) potential of medium sized oligomers in metagenomic analysis and (iii) oligonucleotide signatures in comparative genomics (presented in detail in section 1.2).

Initial results exactly confirmed the findings of Weinel et al. (2002) for *P. putida* and *P. aeruginosa*. 8-14mers were also found to be overrepresented in other Pseudomonads, and later to varying degrees in all sequenced prokaryotes. A distinct contrast was found between those oligomers found to be highly overrepresented, albeit with a search restricted only to the genus *Pseudomonas*, and those which were less overrepresented at a lower chi-squared overrepresentation level across most bacteria. The highly overrepresented oligomers were found to accumulate in either coding or non-coding regions, and if in coding regions, preferred one coding frame in over 80 % of cases across all *Pseudomonas* sp. Many of these are likely to be very common amino acid sequences, or genome wide recognition sequences involved in replication and repair machinery such as chi-sites (Arakawa et al. 2007). In contrast, globally overrepresented oligomers at lower chi-squared levels did not prefer any reading frame when present in coding regions.

Mid-length 8-14 bp oligomers, like the shorter monomers, dinucleotides (Karlin et al. 1997) and tetramers (Reva and Tümmler 2005) before them, were found to be useful in discerning divergent genomic regions. Interestingly, divergent genomic regions can be identified due to their lack of globally overrepresented 8-14mers. We speculate that genome wide processes such as replication and repair, DNA structural biases (Pride et al. 2003), or maintenance of regulatory machinery (Bush and Lahn 2006), the forces which drive maintenance of genomic signatures, slowly facilitate the invasion of overrepresented oligomers into these regions. This would imply that newly integrated genomic islands are free of globally overrepresented oligomers, which are however more obviously present in stably associated islands. Indeed some putative formerly divergent genomic regions, such as an inactive phage in *P. putida*, are almost indistinguishable from the genomic oligomer patterns due to putative pattern amelioration (van Passel et al. 2006, Ganesan et al. 2008). Similar examples can be located and oligomer parameters viewed in the online

SeqWord Genome Browser, another contribution of this thesis (Ganesan et al. 2008).

One of the original use cases of oligomers was intended in the field of metagenomics. To this end the counts and specificity of oligomers was assessed for all species in the genus *Pseudomonas*. Counts of individual highly overrepresented oligonucleotides varied from 59-989 per genome. The oligomers in coding regions were not tightly clustered around specific genomic loci, but rather spread across coding regions of the whole genome. This is optimal for usage in metagenomics, since reads are gained from random loci scattered throughout multiple genomes. Oligonucleotides present in the large *Pseudomonas* sp. genomes were appropriate for metagenomic usage. However, metagenome researchers are not solely interested in Pseudomonads, but in inferring the presence of all sequenced bacteria in their metagenomic datasets. This implies that sufficient distinct overrepresented oligomers must be present in each genome for the principle proved in the genus *Pseudomonas* to function for all bacteria. Harvesting sufficient oligomers is not a problem, but finding database-unique oligonucleotides is not possible. The lack of unique oligomers is apparent when one considers that octamers are most likely to be present frequently enough across the whole sequence to be useful, and the paucity of combinations possible for an octamer (65536). As the number of sequenced genomes continues to grow, this problem will become more intractable, implying either combinations of oligomers or larger oligomer sizes with greater specificity are necessary for metagenomic identification (see section 3 below).

2. Comparative genome analysis

The number of sequenced microbial genomes is growing so rapidly there is a need for simple tools to compare them. This part of the thesis details programs and analytical techniques for whole genome analysis. We demonstrate that compositional approaches such as oligonucleotide usage are useful for both whole genome and gene scale assessment of genomes. Furthermore, results from analysis of Blast based core genome and phylogenetic trees are reported here.

Bacteria have an oligonucleotide signature which is stable over most of the genome except for the accessory regions. These signatures can be analysed by examining patterns in short and mid-length oligonucleotide parameters. Global genomic architecture such as core and accessory genomic regions and location of genomic repeats can then be discerned by the deviation of these oligonucleotides from average. This functions even if only nucleotide sequence is available as input. Thus, the methods displayed here also find application in analysis and quality control of unfinished genomes during the sequencing process. For example, unsequenced or low quality

regions can be quickly recognised by their anomalous oligonucleotide content. Because global genomic 8-14mer repeats do not generally occur in coding regions, annotation quality, particular of low confidence hypothetical proteins, may be assessed by a simple visual scan using OligoCounter's graphical OligoViz application.

Pseudomonas aeruginosa strains have a highly conserved core genome with several insertion sites located in tRNA genes where accessory elements, such as genomic islands, are integrated. Oligomer usage is clearly distinct from the genomic average in the 108 kbp pathogenicity island PAPI-1 in strain PA14. Furthermore, the repetitive nature of *Pseudomonas* sp. chromosomes can be objectively and simply compared with oligonucleotides. If we take two well represented genera of the genus, *P. aeruginosa* and *P. syringae*, it is immediately apparent that *P. syringae* contains many less oligomers than the former according to oligonucleotide usage. Usefully, this fact correlates with the difficulty of building long contigs from short reads *de novo* from these two species, which is relatively easy for *P. syringae* and difficult for *P. aeruginosa* (see section 1.4 below). Another species, the entomopathogenic *P. entomophila*, is dominated by non-coding repeats that are constituents of the known repetitive extragenic palindromic elements. Overrepresented oligomers could be divided into two classes, either non-coding repeats or coding elements that are almost mutually exclusive. It is difficult to test the function of global repeats such as the observed oligomers. However, Chi-sites (5'-GCTGGTGG-3'), which are involved in the repair of double stranded breaks in chromosomal DNA by modulating the activity of an exonuclease and RecBCD helicase (Sourice et al. 1998), were observed in the data.

In the completely unrelated phylum Chlorobi, both tetramers and mid sized oligomers could also distinguish anomalous transposon or phage related sequences from the core genome. For example, *Chlorobium chlorochromatii* CaD3 contains several giant genes of over 100 kbp in size. These fascinating large constructs contain repeated domains which disrupt the oligonucleotide signature characteristic of core genome sequences, since long stretches of repeated amino acids are uncommon in prokaryotic genomes. In contrast, the first sequenced member of the phylum, *Chlorobaculum tepidum* TLS contains very few divergent regions and no giant genes, which is evident at a glance from a oligomer genome atlas. Across all Chlorobi genomes, commonalities were noted in presence of secretion, transport, cell wall and extracellular constituents in anomalous genomic regions. While much interest has centred around the possible role of horizontal gene transfer of photosynthetic biosynthesis genes in the Chlorobi due to their apparently ancestral photosynthetic apparatus (Raymond et al. 2002), we found that most of these genes displayed

oligonucleotide signals more typical of core genomes. Of course, this may simply suggest that photosynthesis associated genes have long been present in the Chlorobi such that their formerly divergent signatures have been ameliorated to better resemble the core genomic average. Thus the evolution of type I photosynthesis reaction centres remains inconclusive but is likely to have developed through a complicated series of events.

In summary, modern computational packages are necessary to provide insight into the abundant data delivered by a genome sequencing project. Efforts must continue to be devoted to this area to further facilitate user friendliness in comparative genomics, as many analyses are still carried out by bioinformatics specialists.

3. *In silico* metagenomic analysis

This section focusses on the development of a novel algorithm, Metatie, which facilitates the use of millions of short reads in metagenomics. Metatie attributes short reads to various taxa with sequenced genomes. The main advantages of this program are compared with existing methods of taxonomic attribution such as rRNA classification.

Ribosomal RNA has been and will remain a very effective tool in qualitatively assessing the species richness of prokaryotic communities in the environment (Tringe and Hugenholtz 2008). However, this method does have obvious limitations. Problems include ribosomal operon copy numbers in bacterial chromosomes vary by an order of magnitude (Klappenbach et al. 2001), RNA reveals little about the functional capacity of a community, and over 99.5 % of randomly sequenced reads remain unused (Biers et al. 2009). We thus developed an approach which allows more specific attributions of short reads to metagenomes, and targets entire genomes, not just single ribosomal RNA genes.

The algorithm described here, Metatie, demonstrates that with two artificial metagenomes, two resequencing projects and three real metagenomes, even 36 bp reads are sufficient to accurately align reads to all sequenced microbial genomes. A recovery rate of about 50 % is demonstrated with a simple artificial metagenome comprising reads from oceanic bacteria with introduced simulated errors. The inherent specificity in the system is sufficient to allow recovery of multiple chromosomes from various multi-replichore strains included in the simulated low complexity test dataset (Mavromatis et al. 2007). 101 of 113 strains (89 %) were correctly detected in this simulated metagenome, with distinctly higher recovery rates to that from the ocean dataset. Errors were on the whole not due to the algorithm, but to the absence of the strain in the reference dataset, and a total

of just four species or strain level false positives were observed. A further validation step included alignment of 6.5 million Illumina reads from *P. aeruginosa* strain CHA to all genomes. Thereafter all *P. aeruginosa* strains were excluded from the reference so as to simulate the effect of having unsequenced related strains in a metagenome, and then all *Pseudomonas* species were excluded. The absolute number of reads attributed declined markedly, which was the desired result for avoidance of false positives. Those reads that were still mapped were assigned to the close relatives of either *P. aeruginosa* or all *Pseudomonas* species. Interestingly, decisively more reads were mapped to one non-Pseudomonad, *Azotobacter vinelandii* DJ, than to other species known to be closely related to the Pseudomonads. These results support the proposal by others (D. Ussery *pers. comm.*), based upon rRNA and whole proteome alignment data, that the newly sequenced *A. vinelandii* belongs to the genus *Pseudomonas*. Only one metagenome sequenced with Illumina 36 bp reads is known to the authors. This incomplete metagenome, derived from human serum in a clinical study of hepatitis B related acute liver failure (Sorber et al. 2008), was heavily contaminated with *H. sapiens* DNA, as expected. After screening and excluding this "contamination", Metatite attributed 635 reads to the hepatitis B virus reference. Only one other virus out of the 3289 present in the reference dataset, Enterobacteria phage M13, was detected in the sample with 48 reads. This fact again demonstrates the specificity of the Metatite approach. Use of the draft genomes newly sequenced in the Human Microbiome project (Turnbaugh et al. 2007) led to novel findings when analysing existing metagenomes. For example, we report the presence of various newly sequenced anaerobes in the Japanese human gut metagenome (Kurokawa et al. 2007). For example, the newly sequenced anaerobe *Gordonibacter pamelaee* was found in significant numbers, and two dominant genera, *Alistipes* and *Akkermansia*, were also detected, despite not being reported in the original early study (Kurokawa et al. 2007).

The Metatite approach has few limitations, aside from an inability to roughly quantify reads from unsequenced prokaryotes. Some other programs, such as Megan, allow taxonomic estimates of these reads to imprecise taxonomic levels, whereas Metatite is designed to accurately distinguish between known species in metagenomic samples. As such, these approaches are complementary. We envisage that the future of comparative metagenomics and metagenomic monitoring will require accuracy in identifying known organisms rather than unspecific guesses at quantifying unknown DNA.

One caveat to metagenomics remains methodologies for accessing environmental DNA. This is naturally a key step in the entire process, which may greatly bias results. For example, some

studies have failed to detect Archaea in metagenomic datasets, in contrast to many other workers (Gill et al. 2006, Kurokawa et al. 2007, Manichanh et al. 2008). A further example is that firmicutes are rarely found in environmental samples (W. Abraham *pers. comm.*). This is due to difficulties in protocols for lysing some particularly durable microbes without introducing some form of systematic errors (Manichanh et al. 2008).

4. Short read assembly

The various novel sequencing technologies all deliver hundreds of thousands to millions or even tens of millions of reads per run. However, they also share at present one characteristic - short reads of the order of 25 - 400 bp. These data are currently used in two main ways. Firstly, with a known related reference sequence reads can be stringently aligned or mapped onto the reference with great accuracy. Typically the large numbers of reads allow a high read coverage of perhaps 20-40 per reference position, so variable base positions such as SNPs and to some extent larger regions can be accurately called. This works very well for eukaryotes, as tens of Illumina short read sequencers are currently used in human resequencing projects at the Sanger Institute in the UK. For bacteria high quality resequencing of two *P. aeruginosa* sublines has been demonstrated in this thesis. Computationally, the reference assembly step is also relatively undemanding. Of course, reference based alignment is only as good as the reference, and a reference must obviously be available. All SNPs, deletions and variable regions are analysed with respect to the reference. Errors in the reference sequence, for which the 99.99 % accuracy of traditional Sanger sequencing quality standards translates to one error per 10000 bp, or 100 per Mb of sequence, appear as SNPs in the sequence under study even if both organisms are identical (Farrer et al. 2009). Finally and most crucially, insertions in the new sequence, i.e. DNA novel to the target organism, are not detected.

De novo assembly without prior information about the sequence is a more challenging step (Pop et al. 2004, Zerbino and Birney 2008). The reads, which in resequencing projects are typically left over after the assembly step, must somehow be grouped by similarity, and then assembled into larger contigs. Computationally, this is much more demanding than the previous step. Reference assembly is processor rather than RAM limited, meaning it will work on a wide range of machines but may take considerable amounts of time. *De novo* assembly is RAM limited as the sequences to be sorted must be stored in memory, and at least several gigabytes of RAM are necessary. Some users have even reported requiring 100-500 gigabytes of RAM when using the Velvet assembler to assemble hundreds of millions of short reads. Contigs generated by short reads tend themselves to

be short. Their length is restricted by repeated DNA sequences, which when larger than the read length necessarily causes the contig extension to be halted. For *P. aeruginosa* we have observed maximum contig sizes of 15-20 kbp across multiple strains, probably due to the abundant repetitive extragenic palindromic (REPs) elements known to be present in this species. Another *Pseudomonas* species, *P. syringae*, has recently been independently resequenced using the same unpaired 36 bp Illumina technology (Farrer et al. 2009, Reinhardt et al. 2009). Using the same assembly algorithm, far longer contigs were created. Even the median length contigs were of the order of 13 kbp, roughly the same size as the maximum length contigs recorded here. The major discrepancies in these results are primarily due to the relative absence of REPs and long oligonucleotide repeats in *P. syringae* (Davenport et al. 2009).

Future algorithms and technologies that are likely to have a big impact on the field include, in addition to increasing read lengths, paired end reads. With this technique two reads are sequenced from DNA around a known spacer distance, which allows better placement by assembly programs, and precise detection of insertions or deletions should the spacer distance be longer or shorter than was expected, respectively. Paired end reads may allow Illumina reads to be used more effectively in *de novo* sequencing, even for whole genomes.

5. References

- Arakawa, K., Uno, R., Nakayama, Y. & Tomita, M. (2007) Validating the significance of genomic properties of Chi sites from the distribution of all octamers in *Escherichia coli*. *Gene* **392**(1-2) 239-246.
- Biers, E. J., Sun, S. & Howard, E. C. (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75**(7) 2221-2229.
- Bush, E. C. & Lahn, B. T. (2006) The evolution of word composition in metazoan promoter sequence. *PLoS Comput Biol* **2**(11) e150.
- Davenport, C. F., Wiehlmann, L., Reva, O. N. & Tümmler, B. (2009) Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ Microbiol* **11**(5) 1092-1104.
- Farrer, R. A., Kemen, E., Jones, J. D. G. & Studholme, D. J. (2009) *De novo* assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol Lett* **291**(1) 103-111.
- Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tümmler, B. & Reva, O. N. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical

regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**, 333.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. & Nelson, K. E. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**(5778) 1355-1359.

Karlin, S., Mrázek, J. & Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12) 3899-3913.

Klappenbach, J. A., Saxman, P. R., Cole, J. R. & Schmidt, T. M. (2001) rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res* **29**(1) 181-184.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**(4) 169-181.

Manichanh, C., Chapple, C. E., Frangeul, L., Gloux, K., Guigo, R. & Dore, J. (2008) A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res* **36**(16) 5180-5188.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., Rigoutsos, I., et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4**(6) 495-500.

van Passel, M., Bart, A., Luyf, A., van Kampen, A. & van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7**(1) 26.

Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. (2004) Comparative genome assembly. *Brief Bioinform* **5**(3) 237-248.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2) 145-158.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**(5598) 1616-1620.

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D. & Dangel, J. L. (2009) *De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res* **19**(2) 294-305.

Reva, O. N. & Tümmeler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.

Sorber, K., Chiu, C., Webster, D., Dimon, M., Ruby, J. G., Hekele, A. & DeRisi, J. L. (2008) The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE* **3**(10) e3495.

Sourice, S., Biaudet, V., Karoui, M. E., Ehrlich, S. D. & Gruss, A. (1998) Identification of the Chi

site of *Haemophilus influenzae* as several sequences related to the *Escherichia coli* Chi site. *Mol Microbiol* **27**(5) 1021-1029.

Tringe, S. G. & Hugenholtz, P. (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**(5) 442-446.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. (2007) The human microbiome project. *Nature* **449**(7164) 804-810.

Weinel, C., Ussery, D. W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C. & Tümmler, B. (2002) Comparative Genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: Orthologs, Codon Usage, Repetitive Extragenic Palindromic Elements, and Oligonucleotide Motif Signatures. *Genome Letters* **4**, 175-187.

Zerbino, D. R. & Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**(5) 821-829.

Abbreviations

bp	-	nucleotide base pair
D	-	the tetranucleotide parameter "Distance", as defined by Reva and Tümmler 2004
GC	-	G+C content, the percentage of summed guanine and cytosine nucleotides in a sequence
PS	-	the tetranucleotide parameter "Pattern skew", as defined by Reva and Tümmler 2004
OUV	-	the tetranucleotide parameter "Oligonucleotide variance", defined by Reva and Tümmler 2004
RAM	-	Random access memory
REP	-	Repetitive Extragenic Palindromic sequence element
rRNA	-	ribosomal ribonucleic acid
SNP	-	single nucleotide polymorphism
tRNA	-	transfer ribonucleic acid

Appendix

1. List of Manuscripts

Part 1

- **Paper** Davenport, C. F., Wiehlmann, L., Reva, O. N. & Tümmmler, B. (2009) Visualization of *Pseudomonas* genomic structure by abundant 8-14mer oligonucleotides. *Environ Microbiol* **11**(5) 1092-1104.

Part 2

- **Paper** Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tümmmler, B. & Reva, O. N.(2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**, 333.
- **Manuscript** Davenport, C. F., Ussery, D.W. & Tümmmler, B. Comparative genomics of Green Sulfur bacteria. Invited submission to *Photosynthesis Research*, August 2009.

Part 3

- **Manuscript** Davenport, C. F., & Tümmmler, B. Abundant oligomers common to most bacteria. In revision at *Plos One*, September 2009.
- **Manuscript** Emmett, W., Davenport, C. F., Labuschagne P., & Reva, O., Oligonucleotide signatures of the bacterial world. In revision at *Environmental Microbiology*, September 2009.

Part 4

- **Manuscript** Klockgether, J., Munder, A., Neugebauer, J., Davenport, C. F., Stanke, F., Larbig, K. D., Heeb, S., Schöck, U., Pohl, T. M., Wiehlmann, L. & Tümmmler B. Genome Diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains. In revision at the *Journal of Bacteriology*, September 2009.

Part 5

- **Manuscript** Davenport, C. F., Langmead, B., Neugebauer, J., Klockgether, J., & Tümmmler B. Metatite - a fast, accurate and adaptable metagenomic classifier for ultra-short reads. Submitted to *Genome Research*, September 2009.
- **Manuscript** Stanke, F., Jansen, S., Davenport, C.F., & Tümmmler, B. Differential decay of Parent-of-Origin Specific Genomic Sharing in Cystic Fibrosis Affected Sibpairs Maps a Paternally Imprinted Locus to 7q34. Submitted to the *European Journal of Human Genetics*, July 2009.
(this manuscript is listed, but not discussed in the text as it reflects work outside the scope of this thesis).

2. Links to programs

Several applications were tested and maintained (SeqWord), or developed (OligoCounter, Metatie) during this thesis:

The SeqWord Genome Browser

<http://genomics1.mh-hannover.de/seqword/>

OligoCounter package with visualisation tools

<http://webhost1.mh-hannover.de/davenport/oligocounter/>

Metatie: a short read metagenomics tool

<http://genomics1.mh-hannover.de/metatie/>

Colin Davenport

Personal Details

Address:	Clinical Research Group, OE 6710, Hanover Medical School D-30625 Hanover, Germany
Telephone:	0049 (0) 511 5326721
Email:	davenport.colin@mh-hannover.de
Nationality:	British/ New Zealand
Date of Birth:	10.09.1980

Education and Qualifications

2006-2009	PhD in Bioinformatics (ongoing) at Hanover Medical School, Germany Three year PhD focussed on occurrence of overrepresented oligonucleotides in prokaryotic genomes. Genomic and metagenomic applications have already been implemented using Java, Python, MySQL, Linux, and R. Some of these programs are available on the internet (OligoCounter, SeqWord). Responsibilities include organising the Pseudomonas 2009 conference, web development, Linux server administration and hosting guest professors.
2005-2006	MRes Master of Research in Bioinformatics (First class marks: 82%) at Exeter University, England. 6 month research project at the Centre for Biocatalysis in the group of Prof. J. Littlechild researching substrate specificity and homology in Baeyer-Villiger monooxygenase enzymes. I used the Modeller homology modelling package, bayesian phylogenetic tools and the Protein Data Bank extensively.
1999-2003	BSc (Hons) in Ecology (First class marks: 72%) at the University of East Anglia, Norwich, England. Final year project was field-based and included complex data manipulation and statistical analyses as well as research writing practice.

Work Experience

2003 - 2005	Research Assistant part time at the Centre for Development Research (ZEF) in Bonn, Germany. Scientific papers were constructively criticised and improved, then edited and internally published, extending written communication skills.
2003 - 2005	Part time assistant in the Finance/Controlling Department at the Caesar Foundation in Bonn, Germany. I upheld the inventory and maintained and assisted in developing Oracle Financials/Discover and Access databases. I also created and maintained the department website using HTML and Excel worksheets with VBA. Good communication skills and ability to work within the office team were critical to success in the position.

Informatics skills:

OS	Linux (since 2004) with Web and Server admin roles, Windows (since 1993)
Databases	MySQL, Oracle with Discoverer, Access
Statistical	R, SPSS
Structural	Modeller, MOE, PyMol, Chimera
Programming	Java (since 2005), knowledge of BioJava, use of Python for scripting, and some experience of Perl and PHP. SQL was used as an analytical tool during my PhD and I have used HTML/CSS since 2004.
Bioinformatics	Maq, Bowtie, Mosaik, JalView, Clustal, MrBayes, Blast, Artemis
General	Office software, GIMP, Photoshop

Skills, Activities and Responsibilities

- Excellent spoken and written German (proven during over 5 years in Germany).
- Enthusiastic catamaran sailor with experience working with disabled sailors at Dartmouth Sailability
- MLTB trained in Mountain Leadership. Regular Walkleader for the Fell Club (1999-2003). First-Aid Certificate. Active mountaineer with experience in the Alps, New Zealand, Norway and Patagonia.
- Active member of bioinformatics.org.

Publications, Oral Presentations and Posters

2009	Publication: Davenport CF, Wiehlmann L, Reva ON, Tümmeler B. Visualization of Pseudomonas genomic structure by abundant 8-14mer oligonucleotides. Environmental Microbiology, Epub Jan 2009.
2009	Invited talk on Metagenomics at Prof. Herzels group at the Humboldt University Berlin, Germany.
2008	Publication: Ganesan H, Rakitianskaia AS, Davenport CF, Tümmeler B, Reva ON. The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. BMC Bioinformatics 9:333.
2008	Oral presentation in the Metagenomics session of the XX International Congress of Genetics, Berlin, Germany.
2007	Oral presentation for the Hanover Biomedical Research School International Advisory Board, Hanover, Germany.
2007	Posters presented at the ISMB / ECCB Conference, Pseudomonas 2009, ProkaGenomics 2007, 1st South African Bioinformatics Conference.

References available on request