Comparative genomics of four *Pseudomonas* species

Lars Juhl Jensen[1,2,3], **Marie Skovgaard**[1], **Thomas Sicheritz-Pontén**[1,4], **Niclas Tue Hansen**[1], **Helle Johansson**[1], **Merete Kjær Jørgensen**[1], **Kristoffer Kiil**[1], **Peter F. Hallin**[1] **and David Ussery**[1]

[1] Center for Biological Sequence Analysis
BioCentrum-DTU
Building 208
The Technical University of Denmark
DK-2800 Lyngby, Denmark

[2] Structural and Computational Biology Unit
European Molecular Biology Laboratory
Meyerhofstrasse 1
D-69117 Heidelberg, Germany

[3] Max-Delbrück-Centre für Molecular Medicine
Robert-Rössle-Strasse 10
D-13092 Berlin, Germany

[4] Department of Medicinal Chemistry
Division of Pharmacognosy BMC
Husargatan 3
Uppsala, Sweden

1 Introduction

The genus *Pseudomonas* is one of the most diverse bacterial genera, containing over 60 validly described species, isolated from sources ranging from plants to contaminated soils and water to human clinical samples. They are obligate aerobic chemoorganotrophs capable of living on a wide range of aliphatic and aromatic carbon compounds. Not surprisingly, the genus *Pseudomonas* is also phylogenetically rather heterogenous, containing several subgroups (3).

The genome sequence of bacteria from four different *Pseudomonas* species has recently been determined. Each of these species represent a major subcluster of the authentic *Pseudomonas* species. *P. aeruginosa* is a ubiquitous environmental bacterium that is one of the top three causes of opportunistic human infections, and was the first species sequenced (30). *Pseudomonas putida* is a versatile saprophytic organism that has attracted considerable attention due to its potential for biotechnological applications (23). The genomes of *P. aeruginosa* and *P. putida* contain many similar genes, but exhibit different codon usage (36). *Pseudomonas syringae* pathovar tomato DC3000 is an organism that is phytopathogenic for tomato and *Arabidopsis thaliana* (6). Finally, *Pseudomonas fluorescens* is a plant growth-promoting rhizobacterium, and the sequence of *P. fluorescens* SBW25 has been finished and is available from the Sanger Center web page [1], although at the time of writing this genome sequence has not been published.

The goal of this study is to compare some of the properties of the genome of these organisms. First, the genomes will be compared at the DNA sequence level. Then gene conservation will be examined, followed by an analysis of sigma 70 factors. Finally, the proteomes will be compared in terms of differences in protein function, based on a keyword analysis. At all levels examined - that is, the DNA sequence, genes encoded, and protein function, *Pseudomonas syringae* appears to stand out amongst the four *Pseudomonas* species.

2 A global view of the *Pseudomonas* genomes

Pseudonomad genomes typically have sizes between 6 and 7 Mbp (see Table 1); when compared to the more than one hundred sequenced bacterial genomes, only a handful of them are larger than the *Pseudomonas* genomes (*e.g.* , the *Bradyrhizobium japonicum* genome of 9.1 Mbp)[2]. *P. aeruginosa* was the largest bacterial genome at that time it was sequenced, although now the genome appears closer to be typical size for an environmental bacteria. Among the *Pseudomonas* genomes sequenced to date, the *P. aeruginosa* genome has the lowest AT-content (33%), whilst the other three strains have 38–42% AT (see Table 1). All four of the *Pseudomonas* genomes contain around 5450 (+/- 100) genes, and between 4 and 7 rRNA operons (see Table 1).

The fraction of repeated sequences, as shown in Figure 1 A, is simple yet important measure of a genome's properties. Global direct and global inverted repeats were calculated as described elsewhere (14), and the fraction of the genome with 80% or more identity to sequences elsewhere in the genome is shown in Figure 1 A. Global repeats include duplicated regions of the chromosome, such as multiple rRNA clusters; the average level of global repeats in bacterial genomes is around 4%. Of the four genomes, only *P. syringae* differs significantly in global repeats from the average of 150 sequenced genomes (solid black bars in Figure 1). The larger amount of global repeats in *P. syringae* are in part reflective of the larger amount of transposable elements in this genome (6).

[1] `http://www.sanger.ac.uk/Projects/P_fluorescens/`

[2] A current listing of sequenced genomes can be found at the GenomeAtlas web resource `http://www.cbs.dtu.dk/services/GenomeAtlas/`

Organism	Size (bp)	% AT	Number Genes	Number rRNA operons
P. aeruginosa PA01	6,264,403	33.4	5566	4
P. fluorescens SBW25	∼6,703,654	40.0	5480	5
P. putida KT2440	6,181,863	38.5	5350	7
P. syringae pv. tomato DC3000	6,397,126	41.6	5471	5

Table 1: Characteristics of four sequenced *Pseudomonas* genomes. The size of *P. fluorescens* SBW25 genome is an estimate based on the current assembly from the Sanger Center.

P. aeruginosa has a higher level of local repeats than the other *Pseudomonas* genomes. This is in part due to the lower %AT content (33%) than the other genomes. As the base composition deviates more from 50% AT content, the chances of finding local repeats increases. Consistant with this is the relatively low local repeat levels of *P. syringae*, which has the %AT content closest to 50% (e.g., 42% AT) of the four species. Local repeats can play a role in DNA mutations, although for GC rich organisms like *Pseudomonas*, the role of local repeats is less clear. Finally, all the *Pseudomonas* genomes have a higher fraction of direct repeats than inverted repeats (both globally and locally); this is a trend seen in most bacterial genomes (14; 1).

Pseudomonas genomes also differ from many other bacterial genomes in that they have a bias towards under-representation of purine stretches, whilst on average most bacterial genomes tend towards on over-representation of purine stretches, as shown in Figure 1 B. In contrast, alternating pyrimidine/purine stretches tend to be over-represented in the *Pseudomonas* genomes. This bias is likely due to enviornmental factors (34; 36), and it is interesting to note that many of the bacteria with fewer purine stretches than expected are soil bacteria [3].

Figure 1: Global DNA properties of *Pseudomonas* genomes. A. Direct and Inverted DNA Repeats. B. Bias in Purine and pyr/pur stretches.

2.1 Genome alignment.

The complete DNA sequences of three *Pseudomonas* genomes were aligned, using the Artimes Comparison Tool (ACT), downloaded from the Sanger Center web site. In Figure 2 the *P. putida* is in the middle, aligned against the *P. aeruginosa* genome (top) and the *P. syringae* genome (bottom). As can be seen from this figure, there are many rearrangements - in particular near the replication origin (e.g., near either edge of the figure), with less sequence conservation near the replication terminus (towards the middle of the figure). Inversions around the origin of replication (often between two rRNA operons) are common in *Pseudomonas* genomes; for example the *Pseudomonas aeruginosa* PAO1 genome sequenced strain is known to have undergone an inversion around the replication origin, relative to other PAO1 isolates (30). Thus, although many of the genes are conserved within the different species, the relative location within the chromosome is likely to be quite different.

[3] Data not shown, but a list of bacterial genomes, sorted by this bias can be found on our web site `http://www.cbs.dtu.dk/services/GenomeAtlas/A_DNA/index_Bacteria_PercentADNA.html`

Figure 2: Alignment at the DNA level of the *P. aeruginosa* (top), *P. putida* (middle) and *P. syringae* (bottom) genomes. Blast hits with a score of 750 or greater are shown.

2.2 Gene comparison

In addition to the published genomes available from GenBank, there are also available several *Pseudomonas* genomes from various strains which are not in one contiguous piece, but likely contain all the proteins. Using BLAST (2), an all-against-all comparison of 7 public available *Pseudomonas* genomes has been constructed. For a systematic comparison, potential genes for each genome was determined using EasyGene(15), with an R-value of 2 as a cut-off. All open reading frames were translated into their amino acid sequence for the BLAST comparison.

A BLAST database was been generated for each proteome and all proteomes were BLASTed against each of the databases. This resulted in a matrix of BLAST reports from which alignments were counted. Only alignments with a minimum of 80% overlapping amino acids and having an espected value (E-value) below $1x10^5$ were included. The parameters were set such that only the best hit per gene was reported.

Based on these results, the table shown in Figure 3 was constructed, which visualises the homology. Each count of homology was calculated relative to the number of genes in each of the genomes. That is, having 400 BLAST hits from a total of 4000 proteins gives 10% homology. For homologies *between* genomes, the highest possible number of homolgy (100%) was illustrated with a dark gray color, whilst the white color represents 60% homology. For homologies *within* genomes, the scale ranges from 0 to 2%. Within-genome homology is observed down the diagonal of the BLAST table, whilst between-homology is observed on each side of the diagonal. Note that the percentage of homology is not identical in A compared with B as when B is compared with A. This is due to the fact that not all genomes have the same number of proteins. That is, on each side of the diagonal, similar hit counts are observed though the percentage will vary slightly due to small differences in gene number.

Inspection of Figure 3 shows that the two *P. aeruginosa* strains are quite similar to each other, both having in common more than 90% of the proteins, as described using the above criteria. However, the two *P. syringae* strains have only about 70% of the proteins in common, and the two *Pseudomonas fluorescens* strains contain less than 60% of the same proteins. Thus based on this limited set of genomes, it appears that the genomic diversity within the *P. aeruginosa* strains are less than that of the other *Pseudomonas* species. An experimental approach has shown that only about 10% of the genes vary in many *P. aeruginosa* strains. Based on the results described here, this number is likely to be much higher for the other *Pseudomonas* genomes. It is also worth noting that the two *P. syringae* strains seem to differ the most from the other genomes, with an average of around 45% homologs; this can be visualised by the lighter colour in the bottom two rows in Figure 3. Thus as with the repeats, the *P. aeruginosa* and *P. syringae* genomes stand out as distinct from the other three, under different circumstances.

Figure 3: Comparison of genes encoded by seven different *Pseudomonas* genomes. See text for details.

2.3 Sequence profile search for ECF type sigma factors

There is great difference in ability of various microorganism to adapt to changes in their environment. At one end of the scale we find parasites and obligate symbiotic lifeforms, which have evolved to live in one very constant environment—*e.g.* inside another cell. *Pseudomonas*

is at the other end of the spectrum, consisting of freeliving species with the ability to adapt to a large number of very different environments. This adaptation to different environment is to a large extent attained through expression of different sets of genes under different circumstances. In bacteria part of this regulation is often controlled by different σ-factors, which can initiate transcription through the recognition of different upstream elements. The number of σ-factors in an organism can be considered one measure of its adaptability. This number varies between 1 for pathogens with small genomes, such as *Mycoplasma genetalium* (12) and 65 for *Streptomyces coelicolor* (4). In the *P. aeruginosa* and *P. syringae* genomes there are 24 σ-factors. A detailed comparison of σ-factors in these two genomes (18) found 13 extracytoplasmic (ECF) σ-factors with homology to *E. coli* FecI, which is involved in iron acquisition.

The σ^{70} related σ-factors are both the most abundant and the most diverse family. The evolutionary unrelated σ^{54}-factors are usually present in a single copy per genome, and have not been included in this analysis, although they are discussed in another chapter in this book, dedicated to sigma factors. The σ^{70} family encompasses the primary and the ECF σ-factors. Little is known about most of the ECF σ-factors, except that many of them are co-transcribed with their negative regulators, known as anti-σ-factors. These are often transmembrane proteins serving and thus thought to act as sensors of the extracellular environment (20).

Starting with a single ECF σ-factor sequence (SigY from *B. subtilis*), PSI-BLAST (2) was used for identifying an initial sequence set of putative ECF σ-factor by searching against a database of all ORFs longer than 300bp from 100 bacterial genomes. A better quality multiple alignment of these sequences was constructed using ClustalW (33) and a profile HMM was constructed using `hmmbuild` from the HMMER package (9). The resulting HMM was used for searching the database again using `hmmsearch` to identify a more accurate ECF set. This procedure, starting from the ClustalW alignment, was iterated until convergence was achieved.

Using this method, we identified σ^{70} related σ-factors in the genomes of 7 *Pseudomonas* species in Figure 3. The simplest possible analysis one can perform is to simply count in each genome the number of genes encoding σ^{70}/ECF transcription factors, and the results are shown in Figure 4. Although for genomes from the first three of the *Pseudomonas* species shown we find at least 23 σ^{70} related σ-factors, for both *P. syringae* genomes, only 13 σ-factors are found. This much lower number of σ-factors in *P. syringae* compared to the other *Pseudomonas* is very surprising as the loss of even a single σ-factor should result in a number of genes no longer being available to the cell as they cannot be transcribed. Thus, once again, *P. syringae* stands out as unique amongst the four species compared.

Figure 4: Comparison of genes encoded by seven different *Pseudomonas* genomes. See text for details.

2.4 Phylogenetic analysis of ECF σ-factors

To analyze which specific groups of σ-factors have been lost in *P. syringae*, we made a tree of all *Pseudomonas* σ^{70}-factors. Using the self consistent profile HMM constructed above, `hmmalign` was used to align all 83 σ^{70}/ECF sequences from the four *Pseudomonas* species to the model. From the resulting alignment, the most conserved columns were extracted, namely those corresponding to match states in the profile HMM. From this core alignment, pairwise evolutionary distances were estimated using the `protdist` program from the Phylip package (10). Finally, an evolutionary tree was constructed using the UPGMA algorithm (Figure 5). Where infinite evolutionary distances were reported by the `protdist` program, a distance of 99 (a very large distance) was used.

All but one of the σ-factors missing in *P. syringae* are of the ECF type; only a small number of these have previously been associated with pathogens and symbionts (21). This fits very well

Figure 5: Phylogenetic tree of 83 *Pseudomonas* σ^{70}-factors, from the genomes in Table 1. Single letter abbreviations are used for each genome: A = *P. aeruginosa*, F = *P. fluorescens*, S = *P. syringae*, and P = *P. putida*.

with *P. syringae* being a plant pathogen, but may at first seem to contradict the high number of ECF σ-factors found in the pathogenic *P. aeruginosa*. However, *P. aeruginosa* is only an opportunistic pathogen and lives freely in soil too.

The loss of FliA, one of the four major members of the σ^{70} family, is more surprising. FliA regulates the expression of the flagellar as well as chemotaxis genes. Since *P. syringae* are known to have flagella (26), they might now be under control of another σ-factor or other regulatory protein. The loss of a σ-factor can either lead to complementation from other systems or loss of function of the genes regulated by the σ-factor (17). While such complementation must have taken place for the flagella, it is likely that some chemotaxis genes have simply been lost. Similarly, the loss of 10 ECF factors is likely to result in loss of function and possibly pseudogenes in the *P. syringae* genome.

The other three major σ-factors (RpoD, RpoH and RpoS) are present in all four *Pseudomonas* species, and form very clear clusters. In all three clusters *P. syringae*, *P. fluorescens* and *P. putida* form a clade with *P. aeruginosa* separating from them earlier in evolution. This is consistent with the 16S rRNA phylogeny. Phylogenies derived from these three σ-factors, however, disagree on the evolutionary relationship between *P. syringae*, *P. fluorescens* and *P. putida*.

3 Systematic functional comparison of complete genomes

Since it is generally not known which sets of genes in a genome are under the regulation of which σ-factors, it is hard to put the above results in a functional context. We approach this problem through a systematic search for differences in the functional content of the genomes.

To accomplish this we extracted all open reading frames longer than 300 bp from each of the four genomes and translated them to their corresponding amino acid sequences. From these BLAST databases were constructed and all pairwise comparisons were performed using gapped BLASTP with low complexity filter enabled (2). Each of these sets was also searched against the SWISS-PROT database (5). The SWISS-PROT database contains only proteins which have been experimentally verified and have a known function. Hence, a good BLAST hit to SWISS-PROT means the chances for knowing the true function of the protein are much greater, since a similar protein has already been characterised in the literature.

Based on these similarity searches, all *Pseudomonas* protein sequences with significant matches (E-value better than 10^{-6}) to SWISS-PROT were identified. For each of these sequences functional information in the form of SWISS-PROT keywords was transfered from the best match. As only sequences with significant matches to SWISS-PROT are used in the subsequent analysis, the large number of random ORFs do not present a problem.

Subsequently, the protein sequences unique to each of the four *Pseudomonas* (but possibly present in other organisms) were identified. An E-value of 10^{-9} was used as cutoff for the BLAST matches for identification of homologs in both the other *Pseudomonas* and in SWISS-PROT. Similarly, the proteins encoded by each genome having homologous genes in all of the three other species were identified, as were the proteins showing homology to each combination of two species.

For each such set of protein sequences, a systematic search for keywords overrepresented, compared to the reference genome, was performed using hypergeometric statistics. The prob-

ability of observing a given keyword exactly x times in a sample is

$$P(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}},$$

where n is the size of the sample, N is the size of the pool (*i.e.* the total number of sequences with significant matches to SWISS-PROT), and m is the number of sequences labeled with the keyword. Using the R statistics package, we have calculated the cumulative probability that each keyword would, at random, occur as many or more times than the actual observed number, *i.e.* the level of significance that a given keyword is overrepresented in a particular partition.

The significance of each keyword also varies depending on which genome is chosen as reference. While this is not a problem when studying proteins present in only one of the *Pseudomonas*, it presents a problem when studying proteins present in several of the genomes.

3.1 Biological interpretation of the keywords

SWISS-PROT keyword	Significance
P. aeruginosa	
Virulence	$3.8 \cdot 10^{-4}$
Plasmid	$8.5 \cdot 10^{-4}$
P. putida	
Methyltransferase	$7.4 \cdot 10^{-5}$
P. syringae	
DNA recombination	$< 10^{-9}$
Transposition	$< 10^{-9}$
Transposable element	$< 10^{-9}$
Plasmid	$4.4 \cdot 10^{-4}$
Tellurium resistance	$4.6 \cdot 10^{-4}$
P. syringae and *P. putida*	
Plasmid	$< 10^{-9}$
Transposable element	$< 10^{-9}$
Hypothetical protein	$< 10^{-9}$
Intron homing	$< 10^{-9}$
RNA-directed DNA polymerase	$< 10^{-9}$
Endonuclease	$9.6 \cdot 10^{-6}$
Multifunctional enzyme	$3.9 \cdot 10^{-5}$
Nuclease	$4.8 \cdot 10^{-5}$
P. syringae and *P. fluorescens*	
Hypothetical protein	$5.5 \cdot 10^{-4}$
P. aeruginosa and *P. putida*	
Outer membrane	$2.1 \cdot 10^{-4}$
Chemotaxis	$5.4 \cdot 10^{-4}$
Fimbria	$8.4 \cdot 10^{-4}$
All but *P. syringae*	
Electron transport	$3.8 \cdot 10^{-5}$
Arabinose catabolism	$1.5 \cdot 10^{-4}$
Arsenical resistance	$4.6 \cdot 10^{-4}$
Flavoprotein	$7.0 \cdot 10^{-4}$
All but *P. aeruginosa*	
Serine protease	$5.1 \cdot 10^{-5}$
Toxin	$1.8 \cdot 10^{-4}$
All but *P. fluorescens*	
DNA-binding	$5.6 \cdot 10^{-4}$
Present in all	
Ligase	$1.8 \cdot 10^{-8}$
Protein biosynthesis	$5.5 \cdot 10^{-7}$
DNA-binding	$2.9 \cdot 10^{-6}$
Transcription regulation	$4.0 \cdot 10^{-6}$
Aminoacyl-tRNA synthetase	$1.2 \cdot 10^{-5}$
Tricarboxylic acid cycle	$1.5 \cdot 10^{-5}$
Ribosomal protein	$5.1 \cdot 10^{-5}$
Helicase	$1.0 \cdot 10^{-4}$
Lyase	$1.6 \cdot 10^{-4}$
Zinc-finger	$1.7 \cdot 10^{-4}$
Metal-binding	$3.1 \cdot 10^{-4}$
Rotamase	$3.4 \cdot 10^{-4}$
Sigma factor	$4.0 \cdot 10^{-4}$
ATP-binding	$8.1 \cdot 10^{-4}$

Table 2: Significantly overrepresented SWISS-PROT keywords associated with differences among the *Pseudomonas* genomes.

Interestingly, the number of significantly overrepresented keywords in each genome suggests *P. syringae* to be functionally most dissimilar to the other *Pseudomonas* species. This agrees well with the much lower number of σ-factors present in its genome, and the fewer genes in common with other *Pseudomonas* genomes in Figure 3.

The three most overrepresented keywords in *P. syringae* are all related to transposons: "DNA recombination", "Transposition", and "Transposable element". Indeed, inspection of the

P. syringae genome reveals the presence of 13 different families of transposons. The keyword "Plasmid", which may appear non-intuitive for chromosomal genes, could indicate that a large number of the genes only found in *P. syringae* may have been incorporated from plasmids.

The final keyword significantly overrepresented among *P. syringae* specific genes, "Tellurite resistance", makes biological sense too, as tellurium resistance has indeed been reported for *P. syringae* (8). The toxicity of tellurite is largely due to it being oxidative. The resistance mechanism involves reduction of the oxidative tellurite to a less harmful compound and subsequent extrusion (32).

In *P. aeruginosa*, the keyword "Virulence" and "Plasmid" are found to be overrepresented among genes only found in this genome. The presence of virulence related proteins should not come as a surprise, as *P. aeruginosa* is known to be an opportunistic pathogen. As was briefly mentioned for *P. syringae*, the presence of proteins typically found on plasmids is also not surprising. In fact, plasmid transfer may very be the mechanism by which *P. aeruginosa* obtained its virulence genes as such genes are often found on plasmids.

Only one SWISS-PROT keyword, "Methyltransferase", displayed significant overrepresentation in the set of proteins only found in *P. putida*. This keyword is closely related to one found to be overrepresented among proteins shared by *P. putida* and *P. aeruginosa*—namely "Chemotaxis". The chemotaxis mechanism allows bacteria to move in an advantageous way by measuring the gradients of repellants or attractants. At the molecular level chemotaxis involves a number of sensory proteins, methyl accepting chemotaxis proteins (MCP), that measure the concentrations of attractants and repellants over time. MCP bind attractants or repellants, either directly or indirectly, through interactions with periplasmic binding proteins. MCP can be methylated and de-methylated by methyltransferases, to alter their activity.

A large number of keywords are associated with proteins common to all four *Pseudomonas*. The vast majority of these would also be expected to be conserved to other species too as they stem from essential processes like transcription, translation, and central pathways of the metabolism (*e.g.* "Tricarboxylic acid cycle"). "Ligase" being the most significant of the keywords can be explained by the involvement of ligases in transcription as well as translation.

The keyword "sigma factor" is one of the more interesting. According to this test the σ-factors are very well conserved in the *Pseudomonas*. Despite *P. syringae* lacking many of the σ-factors found in other *Pseudomonas*, σ-factors are still overrepresented among the proteins conserved in all four *Pseudomonas*.

Among the proteins unique to/absent from each of the four *Pseudomonas*, it is clear that *P. syringae* has both the most significant and largest number of significantly overrepresented keywords. The functionality encoded by the *P. syringae* genome differs from the other *Pseudomonas*—a finding consistent with the study of σ-factors possibly indicating loss of function in the *P. syringae* genome, as well as the BLAST results in Figure 3.

4 Genome comparison atlases

Knowing that functional differences exist among the *Pseudomonas*, it is interesting to examine if these functions are localized within particular regions unique to the strain. Regions that are found in only in or more of the genomes can be detected in a genome comparison atlas.

4.1 Atlas visualization

Methods such as the "atlas visualization" have previously proven very useful for gaining an overview of local variations in many different genomic features (14; 25; 29). We have previously used Genome Atlases to compare the *P. aeruginosa* and *P. syringae* genomes (36). Atlases will be used here for comparing genomic features of the four *Pseudomonas* species, as

well as for correlating their differences to DNA properties and the functional characterizations obtained from the keyword analysis.

The matches between the 3 published *Pseudomonas* genomes were found by aligning all annotated protein coding genes against each other; all open reading frames longer than 300 bp were translated and used for *P. fluorescens*. The BLAST matches are visualized as the outer three circles in Figures 6, 7, and 8, representing each of the organisms against which BLAST searches were performed and a fourth circle representing matches against the SWISS-PROT database (5). For every protein the negative logarithm of the E-value of the most significant match was calculated (imposing a maximum score of 15 for highly significant matches). These values were mapped to the chromosomal location of the corresponding genes. The color scale was reversed so that only lack of hits appear in the circle. This allows us to detect clusters of proteins that are found in only a subset of the analyzed organisms. On each atlas we also show the localization of the most significant keyword (or set of keywords) for the specie in question. These keywords were inferred from BLAST matches to SWISS-PROT and regions were color coded according to their density of genes having the keyword(s).

In addition to these sequence similarity based circles, a circle showing the local AT-content is included since it contains local contextual DNA information along the genome. It is visualized using a double sided color scheme where regions of with unusually high or low AT-content are highlighted in red and cyan, respectively. The AT-content is closely correlated to many DNA structural properties (25), but is also of interest because foreign DNA acquired through lateral gene transfer often has a different AT-content (13). A second circle shows the *position preference* (one of the structural parameters least correlated to AT-content) which is a measure of anisotropic flexibility and regions of low position preference are generally correlated with more highly expressed genes (27; 24).

The AT-content only contains part of the information that can be encoded by the local base composition. The rest can be represented by a class of parameters called *skews*, which have proven very useful for locating the origin and terminus of replication in bacteria (16).

The skews are independent of the AT-content and instead represent preferences for having G's over C's (the GC-skew) or A's over T's (the AT-skew) on one of the two DNA strands. Both the absolute and relative strength of these skews vary greatly between organisms, as does the orientation of the AT-skew compared to the GC-skew. In the case of *Pseudomonas*, the skews are strong and anti-correlated; they can therefore favorably be combined into a single skew, the *keto skew*, defined as $(G + T - A - C)/N$ (19; 14).

4.2 *P. aeruginosa*

Figure 6: Genome comparison atlas of *P. aeruginosa* PA01. 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 legend, and the innermost circle corresponding to the bottom-most feature in the legend. Dark bands in the outer four circles represent the location of genes which are not conserved in other *Pseudomonas* genomes or in SWISS-PROT (blue circle, fourth lane in the atlas). The inner three circles contain DNA structural properties, as described in the "Atlas Visualisation" section of the text.

Figure 6 shows a Genome Comparison atlas for *P. aeruginosa*. The outermost four circles map BLAST hits (or actually the absence thereof) to genes encoded by the *P. aeruginosa* genome. Since so many of the genes are in common for the different genomes, the places where genes are NOT conserved are shaded dark. For example, there are two prophages present in *P. aeruginosa* (between 0.5 M and 1.0 M bp, labelled in Figure 6), which are not found in the other genomes (dark bands in the four outermost circles - note that the first prophage does have

a match (hence no dark band) in the *P. syringae* genome). The inner three circles contain DNA structural properties, as described above.

P. aeruginosa separated from the other *Pseudomonas* species earlier in the course of evolution and has since then aquired functions which allow it to be a oppotunistic pathogen. In spite of this, the majority of the genes in *P. aeruginosa* have homologs in the other *Pseudomonas* genomes, as can be seen from the relatively few dark lines (representing poor or not matches) in the outer circles. For genes in *P. aeruginosa* but not in the other *Pseudomonas* genomes, two SWISS-PROT keywords were found to be significantly overrepresented; "plasmid" and "virulence". Matches to SWISS-PROT and the other *Pseudomonas*, as well as the position as genes related to the keyword "virulence" can be seen in Figure 6. The keyword "plasmid" is not plotted as it is associated with a large number of different functionalities and is thus found throughout the genome.

Unlike the other *Pseudomonas* species in this study, *P. aeruginosa* is known to be a pathogen infecting humans, it should therefore be possible to see regions of genes that are unique to *P. aeruginosa* and have the keyword "virulence". The area marked *Type III secretion* is a good example of that. The type III secretion system is known to be a determinant of the virulence of *P. aeruginosa* (37; 11). The Hrp type III secretion system that is found in *P. syringae* and *P. fluorescens* is capable of eliciting a hypersensitive response in plant cells (7). The similarity between the Hrp secretion system and the type III secretion system seen in *P. aeruginosa* is, however, only weak as they serve different functions.

Two prophages are also labeled on the atlas. One, prophage Pf1, is only found in *P. aeruginosa*, and a putative bacteriophage marked 100kb upstream appears to have a homolog in *P. syringae* but not in the two other *Pseudomonas*. The two labeled prophage regions have very high AT-content compared to the rest of the genome as is the case for a region at 4.7Mb. This region was—because of the difference in AT-content—suggested to be a candidate of horizontal gene transfer (30). The very low value of position preference also seen in this region indicates that the region is likely to contain highly expressed genes. The corresponding proteins match sequences in SWISS-PROT and the region as such is conserved among *Pseudomonas*. The region is an ancient cluster of genes encoding several ribosomal proteins, subunits of the RNA polymerase, and SecY (31). In *P. syringae* and *P. putida* the cluster is located at approximately 0.5 Mb, indicating that the cluster was relocated as part of the proposed chromosomal inversion around the origin (30).

4.3 *P. putida*

Figure 7: Genome comparison atlas *P. putida* KT2440. The circles are as described in Figure 6.

In *P. putida* one keyword was found to be significantly overrepresented: "Methyltransferase". Unlike the virulence genes in *P. aeruginosa*, these genes are not found to form clusters within the genome of *P. putida*. A number of other regions found only in *P. putida* are seen in Figure 7. However, as most of these acquired regions contain genes that are not found in SWISS-PROT either, one can only speculate what their functions are. In a few cases where the functionality of some of the genes is known, they have been found to increase the metabolic proficiency of *P. putida* (35; 23).

Some of the regions that have matches to SWISS-PROT correspond to *prophages*; these were also found to be connected to viral proteins by the phylome analysis. The *T7 prophage* stands out in particular by also having low position preference and inverted keto skew compared to its surroundings. Incidentally, the only other region with strong skew inversion present in the *P. putida* genome also has very low position preference. It corresponds to a gene encoding a

putative 8682 aa surface adhesion protein, making it the largest of the annotated protein coding genes in *P. putida*.

Many of the inserted regions in *P. putida* (but not the prophages) have a very high AT-content compared to the genome average. This, together with the number of insertions, indicates that *P. putida* is highly suspectible to gene transfer (35). The region labeled *Thymidylate insertion* is an example of one such region; it is a 60kb element inserted in the thymidylate kinase gene. While most of the genes in this region are of unknown function, four adjacent genes have by homology been associated with arsenic/arsenite resistance. In contrast to *P. aeruginosa*, *P. putida* is known to be resistant to both arsenite and arsenate (22).

4.4 *P. syringae*

Figure 8: Genome comparison atlas of *P. syringae* pv. tomato DC3000.

While *P. putida* is known to be resistant to arsenic containing compounds, the keyword "tellurium resistance" was found to be overrepresented among *P. syringae* specific genes. The genes with this keyword are localized as one gene cluster, which is labeled in Figure 8.

From the large number of significantly overrepresented keywords found for *P. syringae*, it is clear that it is very different from the other *Pseudomonas*. The most significant keywords were "DNA recombination", "transposable element" and "transposition", which are closely related and thus shown together as one circle (black circle in Figure 8).

In addition to the genes corresponding to these keywords, a number of genes exist that are found neither in other *Pseudomonas* genomes nor in SWISS-PROT. Similar to what was observed in *P. putida*, these genes are clustered within the genome, although the number of clusters is larger in *P. syringae*. Once again, one can only speculate about function of these genes.

Two such unknown regions of approximately 100kb each are found near the *origin* and *terminus* of replication. This pattern is neither seen in *P. putida* nor in *P. aeruginosa*. The positioning of unknown genes near the origin is surprising as this region often contains highly expressed household genes. It is thus intriguing that the region has low position preference (indicating the possibility of highly expressed genes), although it should be noted that this is the case for many other inserted regions as well.

Next to the terminus is located a cluster of genes that are present in *P. putida* too but not in the two other *Pseudomonas* genomes. A number of genes in this region further have similarity to SWISS-PROT sequences from the nitrogen fixating soil bacterium *Rhizobium sp.* strain NGR234. However, the function of these is not known, explaining why this group of genes was not discovered in the keyword analysis. Given that *Rhizobium* lives in symbiosis with plants while *P. syringae* is a plant pathogen, we suggest that this region is a likely candidate of lateral gene transfer. This is supported by the high AT-content of the region compared to the genome average as well as by the region being flanked by transposons

Figure 9: Zoom of a region with many transposable elements and genes unique to *P. syringae*.

Although transposable elements are dispersed throughout the genome, they appear to be especially frequent within the otherwise uncharacterized regions. This is in particular true for the region shown in Figure 9 (this region is labeled *Zoom* in Figure 8). From the zoom it is obvious that the subregions of unknown genes are flanked by genes associated with transposition. The region does not only contain genes of unknown function—genes involved in the synthesis

of the polyketide coronafacic acid are also located in the region. Futhermore phylome analysis suggests a possible source for these genes as they are most closely related to actinobacterial genes. Another observation made from phylome analysis was the relatively high number of *P. syringae* proteins predicted to be of eukaryotic origin, which is possibly due to lateral gene transfer from the host organism.

5 Conclusions

We have compared the genomes of four different *Pseudomonas* species, using several different approaches, and find that, depending on the features examined, in some ways *P. aeruginosa* stands out as unique amongst the four, but in many other ways, in particular in genome variation, *P. syringae* is significantly different from the other three. It is the fundamental idea of comparative genomics to use the entire genomes as the basis for comparison—however suprisingly few methods have been developed for actually doing it. We present a new approach: a systematic statistical search for differences between the gene functions present in a set of genomes. This allows us to, without prior knowledge, obtain hints about biological properties unique to each organism.

When it comes to gaining an overview of complete genomes, it is difficult to overestimate the importance of good visualization techniques. They enable scientists to discover relations in the data that would otherwise go unnoticed. In the present chapter we have used such a method, the atlas visulization, to both compare the four *Pseudomonas* genomes to each other and relate their differences to DNA structural properties as well as gene functions.

Despite strong phylogenetic support for *P. aeruginosa* being evolutionary separated from the other three *Pseudomonas*, it is clearly *P. syringae* that stands out when one performs the comparison at the genomic level. This suggests that the genome of *P. syringae* must have undergone large changes within a relatively short timespan as part of adapting to a symbiotic/parasitic lifestyle. Many such changes in the *P. syringae* genome are revealed by our analysis, include the loss of many genes and uptake of others, which is likely to be mediated by the large number of transposons dispersed throughout its genome. The same methods, when applied to the *P. aeruginosa* genome, highlighted a number of features related to its pathogenicity, in particular the type III secretory system.

6 Acknowledgements

This work was supported by grants from the Danish National Research Foundation and the Danish Natural Science Research Council. Lars Juhl Jensen is funded by the Bundesministerium für Forschung und Bildung, BMBF-01-GG-9817. Marie Skovgaard is funded by EU Cell Factory Project, Screen, QLK3-CT-2000-00649. Thomas Sicheritz-Pontén is funded by Knut and Alice Wallenberg's Foundation.

Preliminary sequence data for the *Pseudomonas fluorescens* strain PfO1 was obtained from the Joint Genomes Initiative web page (`http://spider.jgi-psf.org/JGI_microbial/html/pseudomonas/pseudo_homepage.html`).

References

[1] Achaz, G., Rocha, E. P., Netter, P. and Coissac, E., 2002, Origin and fate of repeats in bacteria. *Nucl. Acids Res.*, 30:2987–2994.

[2] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J., 1997, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.*, 25:3389–3402.

[3] Anzai, Y., Kim, H., Park, J., Wakabayashi, H. and Oyaizu, H., 2000, Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. *Int. J. Sys. Evolut. Microbiol.*, 50:1563–1589.

[4] Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quali, A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.-H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.-A., Rutherford, K., Rutter, S., Seeger, K., Saunder, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, K., Barrell, B. G., Parkhill, K. and Hopwood, D. A., 2002, Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417:141–147.

[5] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M., 2003, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.*, 31:365–370.

[6] Buell, C., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I., Gwinn, M., Dodson, R., Deboy, A., R.T.and Durkin, Kolonay, J., Madupu, R., Daugherty, S., Brinkac, L., Beanan, M., Haft, D., Nelson, W., Davidsen, T., Zafar, N., Zhou, L., Liu, J., Yuan, Q., Khouri, H., Fedorova, N., Tran, B., Russell, D., Berry, K., Utterback, T., Van Aken, S., Feldblyum, T., D'Ascenzo, M., Deng, W., Ramos, A., Alfano, J., Cartinhouri, S., Chatterjee, A., Delaney, T., Lazarowitz, S., Martin, G., Schneideri, D., Tang, X., Bender, C., White, O., Fraser, C. and Collmer, A., 2003, The complete genome sequence of the arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomatodc3000. *Proc. Natl. Acad. Sci. USA*, 100:10181–10186.

[7] Collmer, A., Badel, J. L., Charkowski, A. O., Deng, W. L., Fouts, D. E., Ramos, A. R., Rehm, A. H., Anderson, D. M., Schneewind, O., van Dijk, K. and Alfano, J. R., 2000, *Pseudomonas syringae* Hrp type III secretion system and effector proteins. *Proc Natl Acad Sci U.S.A.*, 97:8770–8777.

[8] Cournoyer, B., Watanabe, S. and Vivian, A., 1998, A tellurite-resistance genetic determinant from phytopathogenic *Pseudomonads* encodes a thiopurine methyltransferase: evidence of a widely-conserved family of methyltransferases. *Biochim. Biophys. Acta*, 1397:161–168.

[9] Eddy, S. R., 1998, Profile hidden markov models. *Bioinformatics*, 14:755–763.

[10] Felsenstein, J., 1989, Phylip—phylogeny inference package (version 3.2). *Cladistics*, 5:164–166.

[11] Francis, M. S., Wolf-Watz, H. and Forsberg, A., 2002, Regulation of type III secretion systems. *Curr. Opin. Microbiol.*, 5:166–172.

[12] Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleishmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek,

D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Petersen, S. N., Smith, H. O., Hutchison, C. A. and Venter, J. C., 1995, The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270:397–403.

[13] Friis, C., Jensen, L. J. and Ussery, D. W., 2000, Visualization of pathogenicity regions in bacteria. *Genetica*, 108:47–51.

[14] Jensen, L. J., Friis, C. and Ussery, D. W., 1999, Three views of microbial genomes. *Res. Microbiol.*, 150:773–777.

[15] Larsen, T. and Krogh, A., 2003, Easygene - a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, 4:21.

[16] Lobry, J. R., 1996, Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, 13:660–665.

[17] Madan Babu, M., 2003, Did the loss of sigma factors initiate pseudogene accumulation in *M. leprae*? *Trends Microbiol.*, 11:59–61.

[18] Martinez-Bueno, M., Tobes, R., Rey, M. and Ramos, J., 2002, Detection of multiple extracytoplasmic function (ecf) sigma factors in the genome of *Pseudomonas putida* KT2440 and their counterparts in *Pseudomonas aeruginosa* PAO1. *Environ. Microbiol.*, 4:842–855.

[19] McLean, M. J., Wolfe, K. H. and Devine, K. M., 1998, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.*, 47:691–696.

[20] Missiakas, D. and Raina, S., 1998, The extracytoplasmic function sigma factors: role and regulation. *Mol. Microbiol.*, 28:1059–1066.

[21] Mittenhuber, G., 2002, An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes. *J. Mol. Microbiol. Biotechnol.*, 4:77–91.

[22] Mynbaeva, B. N., Okorokov, L. A., Abdrashitova, S. A. and Ilialetdinov, A. N., 1984, Phosphate and glucose accumulation by *Pseudomonas* cultures in relation to their arsenic resistance. *Mikrobiologiia*, 53:822–825.

[23] Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Martins dos Santos, V. A., Fouts, D. E., Gill, S. R., Pop, M., Holmes, M., Brinkac, L., Beanan, M., DeBoy, R. T., Daugherty, S., Kolonay, J., Madupu, R., Nelson, W., White, O., Peterson, J., Khouri, H., Hance, I., Chris Lee, P., Holtzapple, E., Scanlan, D., Tran, K., Moazzez, A., Utterback, T., Rizzo, M., Lee, K., Kosack, D., Moestl, D., Wedler, H., Lauber, J., Stjepandic, D., Hoheisel, J., Straetz, M., Heim, S., Kiewitz, C., Eisen, J., Timmis, K. N., Dusterhoft, A., Tummler, B. and Fraser, C. M., 2002, Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ. Microbiol.*, 4:799–808.

[24] Pedersen, A. G., Baldi, P., Chauvin, Y. and Brunak, S., 1998, DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, 281:663–673.

[25] Pedersen, A. G., Jensen, L. J., Stærfeldt, H. H., Brunak, S. and Ussery, D. W., 2000, A DNA structural atlas of *E. coli*. *J. Mol. Biol.*, 299:907–930.

[26] Roine, E., Wei, W., Yuan, J., Nurmiaho-Lassila, E. L., Kalkkinen, N., Romantschuk, M. and He, S. Y., 1997, Hrp pilus: an hrp-dependent bacterial surface appendage produced by *Pseudomonas syringae* pv. tomato DC3000. *Proc Natl Acad Sci U.S.A.*, 94:3459–3464.

[27] Satchwell, S. C., Drew, H. R. and Travers, A. A., 1986, Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, 191:659–675.

[28] Sicheritz-Ponten, T. and Andersson, S. G., 2001, A phylogenomic approach to microbial evolution. *Nucl. Acids Res.*, 29:545–552.

[29] Skovgaard, M., Jensen, L. J., Friis, C., Stærfeldt, H.-H., Worning, P., Brunak, S. and Ussery, D., 2002, The atlas visualisation of genome-wide information. In B. Wren and N. Dorrell (eds.), Methods in Microbiology, volume 33, pp. 49–63. Academic Press, London, UK.

[30] Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K., Wu, Z. and Paulsen, I. T., 2000, The complete genome of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature*, 406:959–964.

[31] Sugita, M., Sugishita, H., Fujishiro, T., Tsuboi, M., Sugita, C., Endo, T. and Sugiura, M., 1997, Organization of a large gene cluster encoding ribosomal proteins in the cyanobacterium *Synechococcus* sp. strain PCC 6301: comparison of gene clusters among cyanobacteria, eubacteria and chloroplast genomes. *Gene*, 195:73–79.

[32] Taylor, D. E., 1999, Bacterial tellurite resistance. *Trends in Microbiology*, 7:111–115.

[33] Thompson, J. D., Higgins, D. G. and Gibson, T. J., 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680.

[34] Ussery, D. W., Soumpasis, D. M., Brunak, S., Stærfeldt, H. H., Worning, P. and Krogh, A., 2002, Bias of purine stretches in sequenced genomes. *Computers in Chemistry*, 26:531–541.

[35] Weinel, C., Nelson, K. E. and Tummler, B., 2002, Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ. Microbiol.*, 4:809–818.

[36] Weinel, C., Ussery, D., Ohlsson, H., Ponten, T., Kiewitz, C. and Tummler, B., 2002, Comparative genomics of *Pseudomonas aeruginosa* PAO1 and KT2440: Orthologs, codon usage, repetitive extragenic palindromic elements, and oligonucleotide motif signatures. *Genome Letters*, 1:175–187.

[37] Yahr, T. L., Goranson, J. and Frank, D. W., 1996, Exoenzyme S of *Pseudomonas aeruginosa* is secreted by a type III pathway. *Mol. Microbiol.*, 22:991–1003.

Figure 1A. Repeats in Pseudomonad genomes

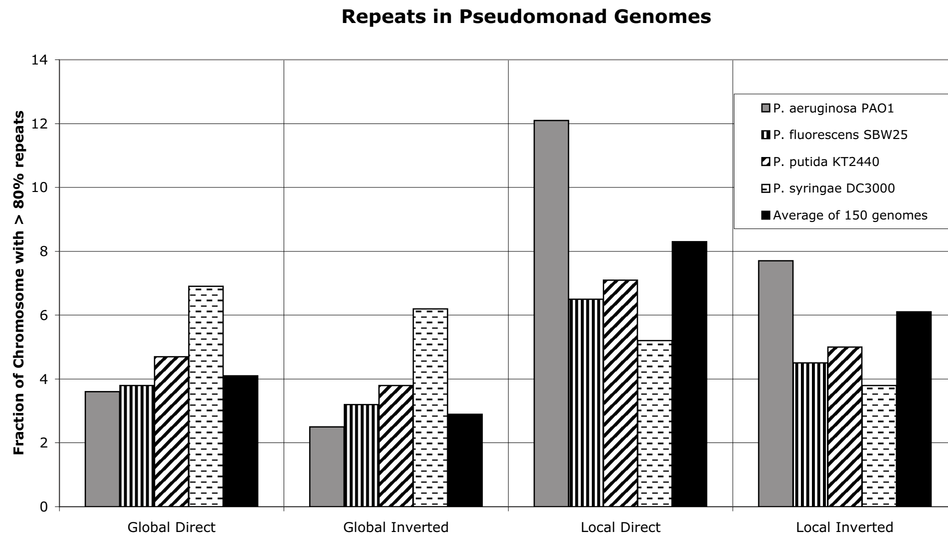


Figure 1B. Bias in purine stretches in Pseudomonad genomes

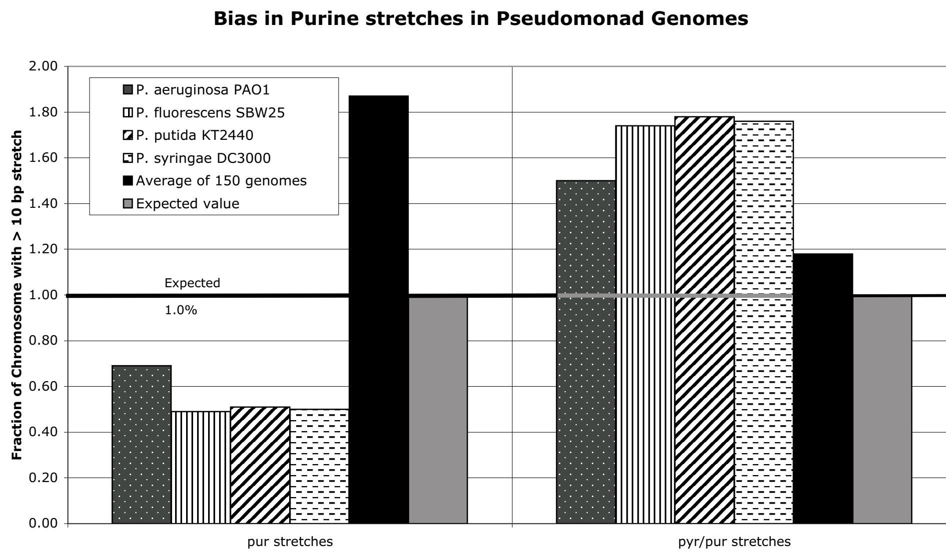
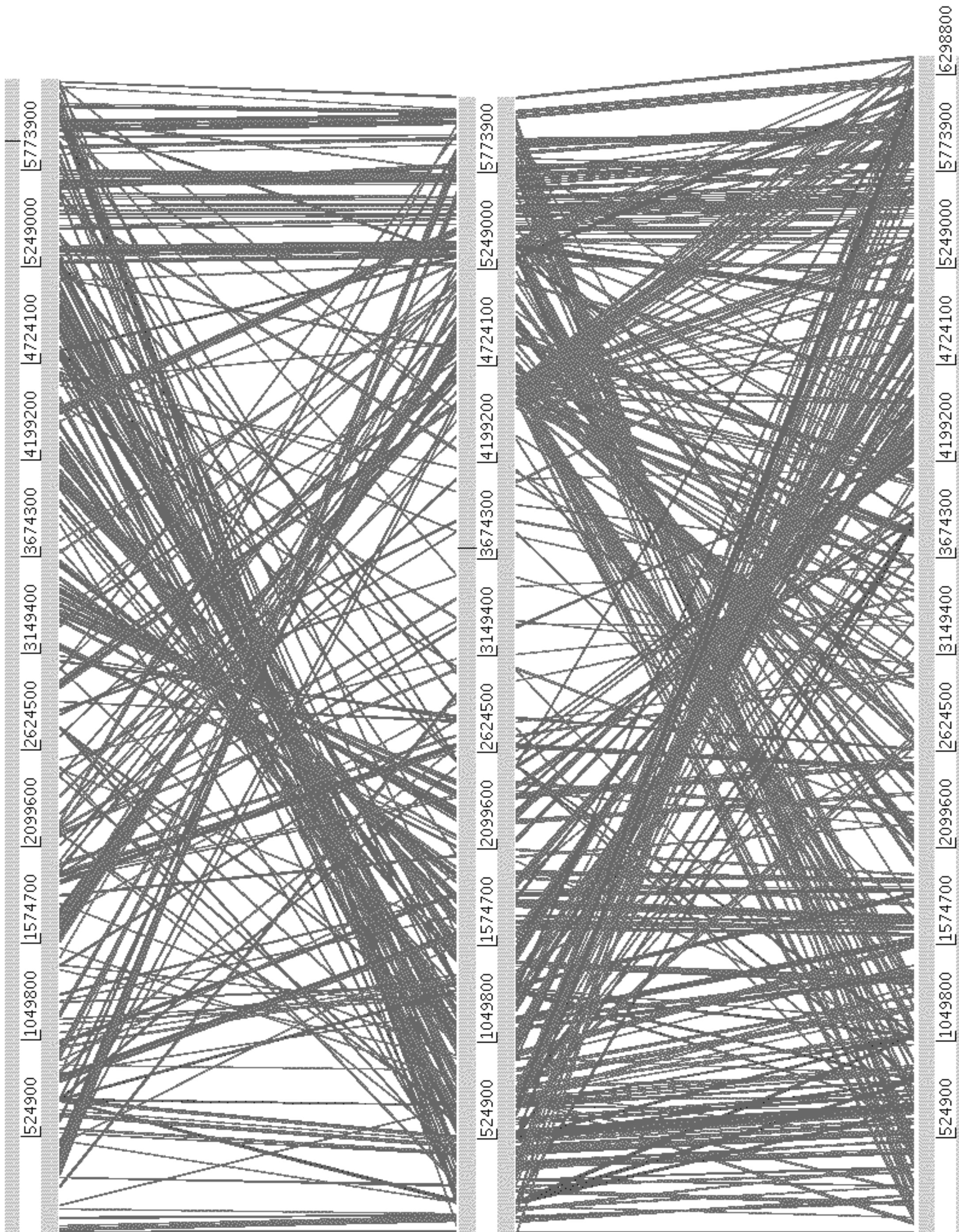
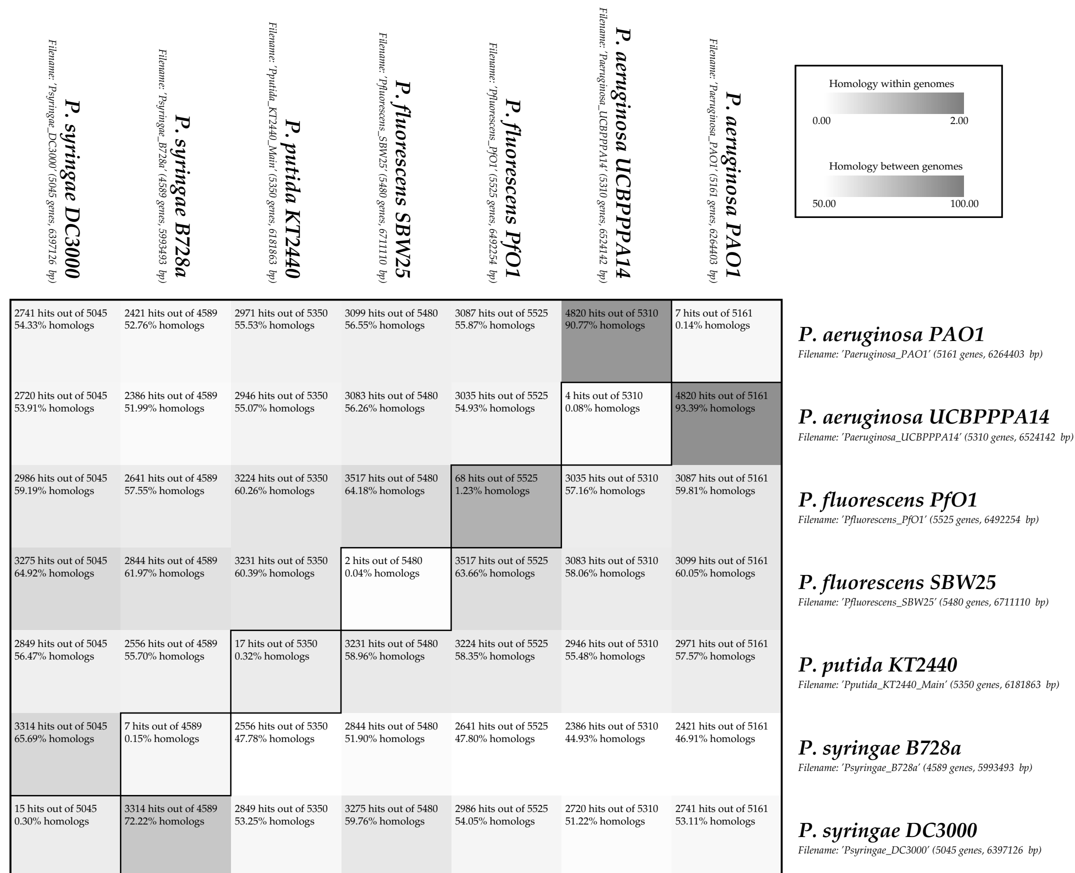


Figure 2



Comparison of Pseudomonas genomes

ALR=80%, E-value < 1x10-5



	P. syringae DC3000 Filename: 'Psyringae_DC3000' (5045 genes, 6397126 bp)	*P. syringae B728a* Filename: 'Psyringae_B728a' (4589 genes, 5993493 bp)	*P. putida KT2440* Filename: 'Pputida_KT2440_Main' (5350 genes, 6181863 bp)	*P. fluorescens SBW25* Filename: 'Pfluorescens_SBW25' (5480 genes, 6711110 bp)	*P. fluorescens PfO1* Filename: 'Pfluorescens_PfO1' (5525 genes, 6492254 bp)	*P. aeruginosa UCBPPPA14* Filename: 'Paeruginosa_UCBPPPA14' (5310 genes, 6524142 bp)	*P. aeruginosa PAO1* Filename: 'Paeruginosa_PAO1' (5161 genes, 6264403 bp)
P. aeruginosa PAO1 Filename: 'Paeruginosa_PAO1' (5161 genes, 6264403 bp)	2741 hits out of 5045 54.33% homologs	2421 hits out of 4589 52.76% homologs	2971 hits out of 5350 55.53% homologs	3099 hits out of 5480 56.55% homologs	3087 hits out of 5525 55.87% homologs	4820 hits out of 5310 90.77% homologs	7 hits out of 5161 0.14% homologs
P. aeruginosa UCBPPPA14 Filename: 'Paeruginosa_UCBPPPA14' (5310 genes, 6524142 bp)	2720 hits out of 5045 53.91% homologs	2386 hits out of 4589 51.99% homologs	2946 hits out of 5350 55.07% homologs	3083 hits out of 5480 56.26% homologs	3035 hits out of 5525 54.93% homologs	4 hits out of 5310 0.08% homologs	4820 hits out of 5161 93.39% homologs
P. fluorescens PfO1 Filename: 'Pfluorescens_PfO1' (5525 genes, 6492254 bp)	2986 hits out of 5045 59.19% homologs	2641 hits out of 4589 57.55% homologs	3224 hits out of 5350 60.26% homologs	3517 hits out of 5480 64.18% homologs	68 hits out of 5525 1.23% homologs	3035 hits out of 5310 57.16% homologs	3087 hits out of 5161 59.81% homologs
P. fluorescens SBW25 Filename: 'Pfluorescens_SBW25' (5480 genes, 6711110 bp)	3275 hits out of 5045 64.92% homologs	2844 hits out of 4589 61.97% homologs	3231 hits out of 5350 60.39% homologs	2 hits out of 5480 0.04% homologs	3517 hits out of 5525 63.66% homologs	3083 hits out of 5310 58.06% homologs	3099 hits out of 5161 60.05% homologs
P. putida KT2440 Filename: 'Pputida_KT2440_Main' (5350 genes, 6181863 bp)	2849 hits out of 5045 56.47% homologs	2556 hits out of 4589 55.70% homologs	17 hits out of 5350 0.32% homologs	3231 hits out of 5480 58.96% homologs	3224 hits out of 5525 58.35% homologs	2946 hits out of 5310 55.48% homologs	2971 hits out of 5161 57.57% homologs
P. syringae B728a Filename: 'Psyringae_B728a' (4589 genes, 5993493 bp)	3314 hits out of 5045 65.69% homologs	7 hits out of 4589 0.15% homologs	2556 hits out of 5350 47.78% homologs	2844 hits out of 5480 51.90% homologs	2641 hits out of 5525 47.80% homologs	2386 hits out of 5310 44.93% homologs	2421 hits out of 5161 46.91% homologs
P. syringae DC3000 Filename: 'Psyringae_DC3000' (5045 genes, 6397126 bp)	15 hits out of 5045 0.30% homologs	3314 hits out of 4589 72.22% homologs	2849 hits out of 5350 53.25% homologs	3275 hits out of 5480 59.76% homologs	2986 hits out of 5525 54.05% homologs	2720 hits out of 5310 51.22% homologs	2741 hits out of 5161 53.11% homologs

Figure 3

Figure 4

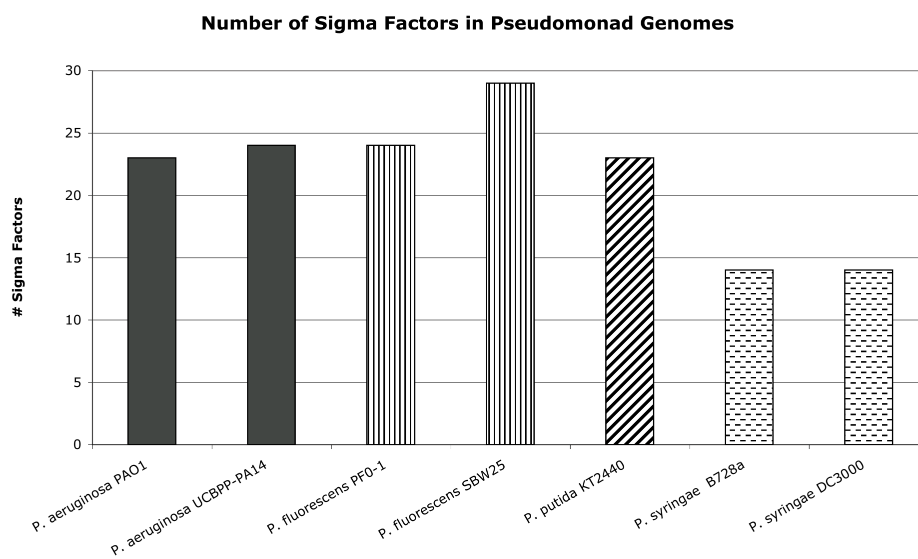
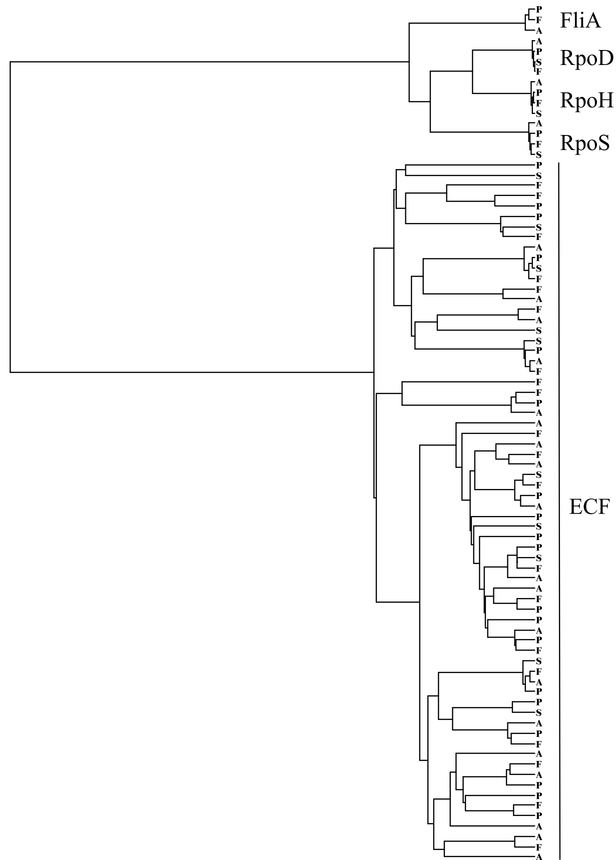
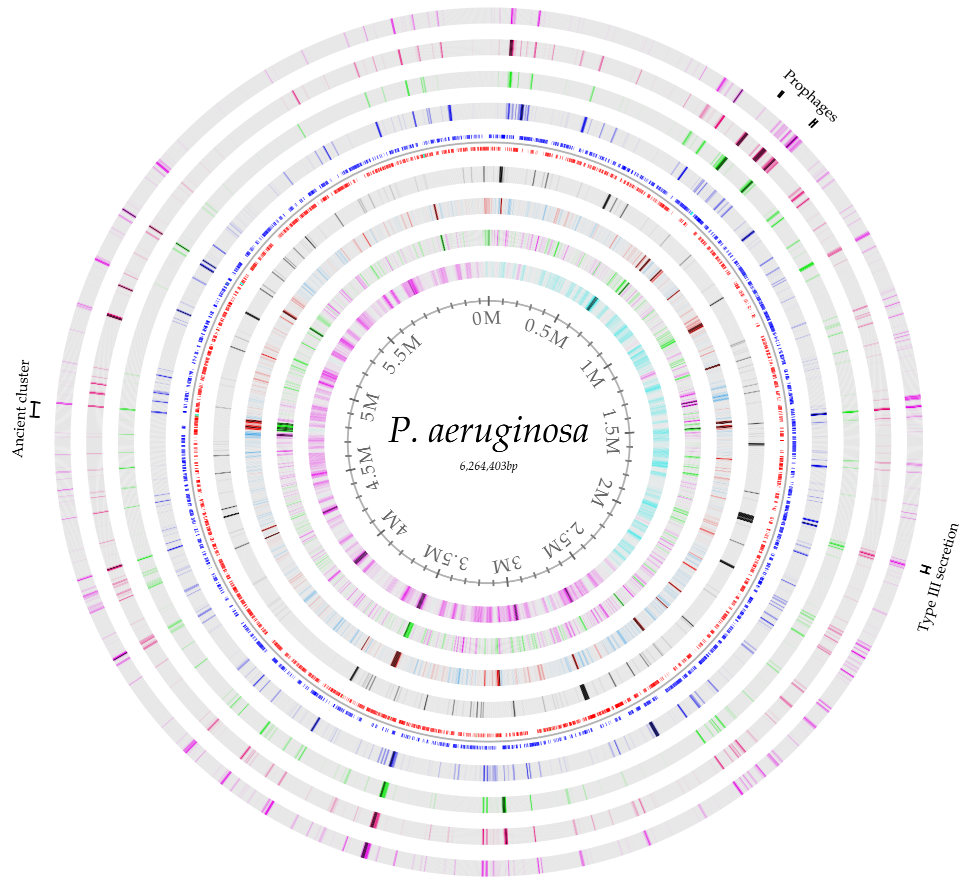


Figure 5





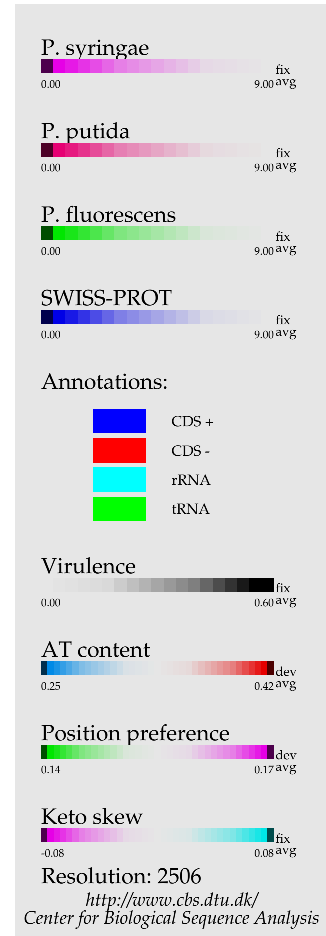
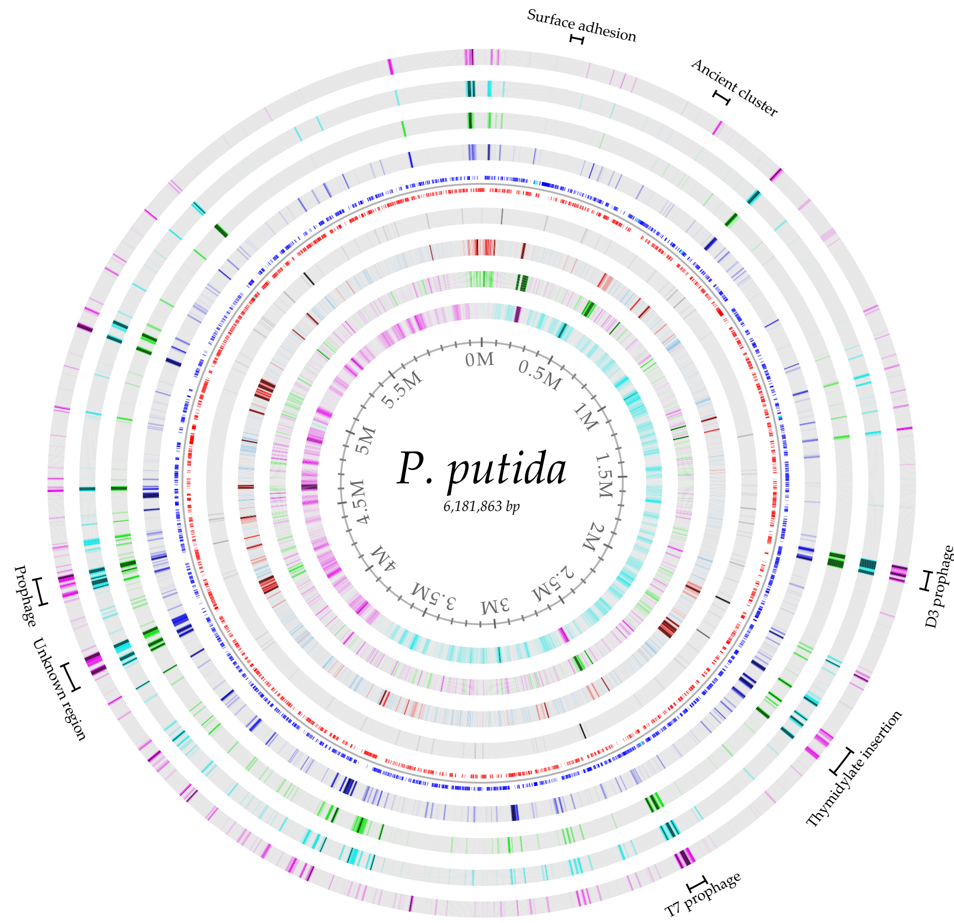


Figure 6



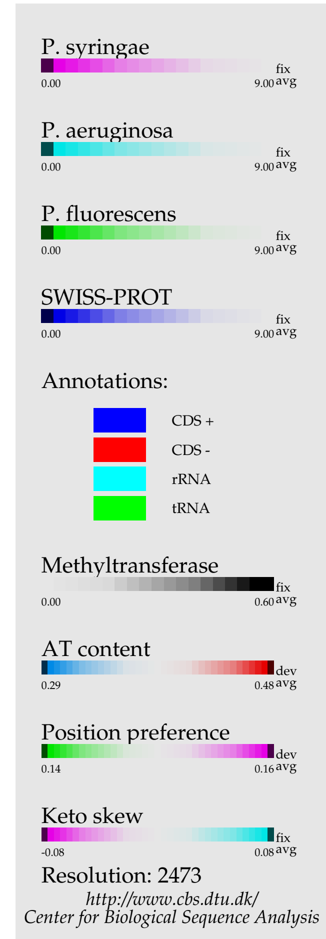
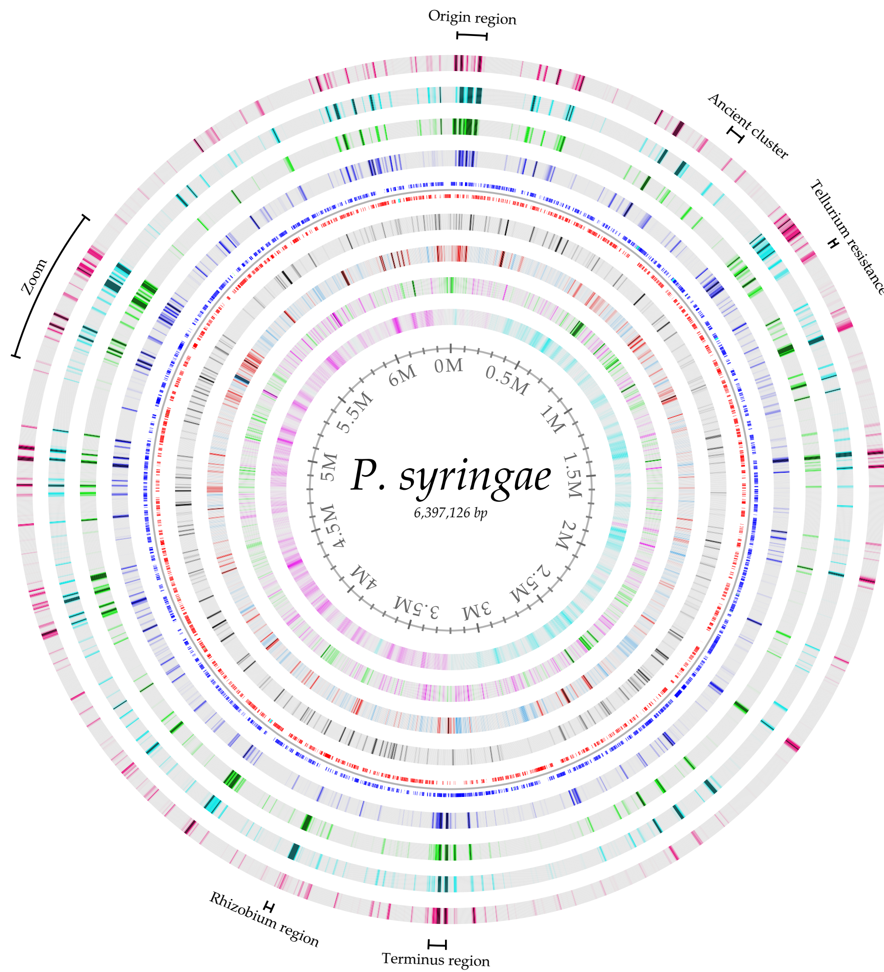


Figure 7



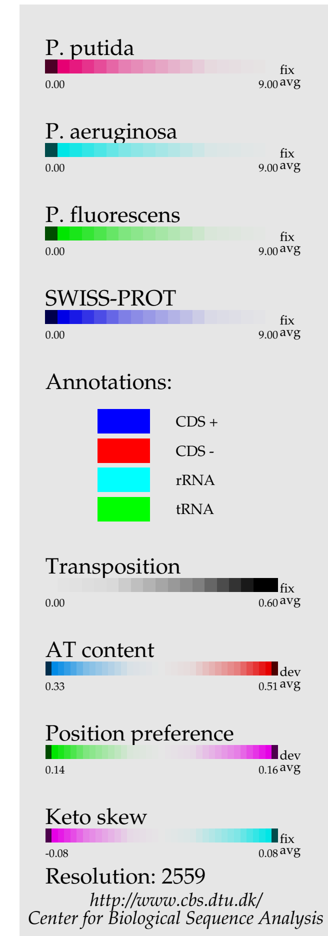


Figure 8

6.1 Figure 9