Whole genome sequence comparisons in taxonomy

Rainer Borriss[1], Christian Rueckert[2], Jochen Blom[2], Oliver Bezuidt[3], Oleg Reva[3]

[1]ABiTEP GmbH, Glienicker Weg 185, D - 12489 Berlin, Germany, [2]Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstraße 27, D-33615 Bielefeld, Germany [3]University of Pretoria, Dep. Biochemistry, Bioinformatics and Computational Biology Unit, Hillcrest, Lynnwood Rd., Pretoria 0002, South Africa.

rainer.borriss@rz.hu-berlin.de; Christian.Rueckert@CeBiTec.Uni-Bielefeld.DE; jblom@cebitec.uni-bielefeld.de; oleg.reva@up.ac.za

Contents

I. Introduction

This chapter is devoted to application of whole genome sequence comparisons in taxonomy. Driven by the rapid progress in sequencing technologies, “low budget” bacterial genomes become increasingly available in a nearly unlimited number. During finalizing this chapter, completed genomes representing 1,604 bacterial and 85 archaeal species were present in the public data bank (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi) reflecting the enormous progress made within sequencing microbial genomes in the last years. With the advent of next generation sequencing, whole genome sequence comparisons will be more and more important for taxonomy, especially valuable in elucidating relationship of groups of closely related bacterial strains which might form a single taxon, a subspecies or just an ecovar within a given species. The aim of this chapter is to hand out a tool set for applying genomics to the interested taxonomist. Using these tools might prove as being useful especially in refining groups of closely related strains, which are not resolved by their 16S rRNA sequence. Here, we will exemplify this approach by selecting a specific group of plant – associated *Bacillus amyloliquefaciens* strains with plant growth promoting properties. In recent years, those strains were increasingly applied as biological substitutes of agrochemicals, mainly used as biofertilizer and for biocontrol of phytopathogenic microorganisms, and nematodes (Chen *et al*., 2007).

Despite the enormous progress made in microbial whole genome sequencing in recent years, the minimal standards in describing of new prokaryotic taxa are founded mainly of a set of microscopic and macroscopic features as cell and colony morphology, physiological and biochemical characters, profiles of fatty acid and cell wall constituents. In addition, whole 16S rRNA gene sequence analysis and, in case of closely related species, DNA-DNA hybridization is also recommended (Logan *et al*., 2009). Unfortunately, those standards are not completely sufficient to discriminate in a satisfying manner closely related taxa, as found with members of the *Bacillus subtilis* species complex. For many years, it has been recognized that these species can´t be discriminated alone on the basis of phenotypic characteristics and 16S rRNA gene nucleotide sequence. Besides fatty acid profiles, that do not yield satisfying results for discriminating closely related bacterial groups, phylogenetic analysis of multiple protein-coding loci has been used as a complementary approach to detect and differentiate novel *Bacillus* taxa (Gatson *et al.*, 2006, Rooney *et al*., 2009). We have successfully used the same approach to discriminate a group of plant-associated *Bacillus* strains related to the *B. amyloliquefaciens* type strain DSM7 and *B. subtilis* 168. Two ecovars, consisting of plant-associated and non-plant associated *B. amyloliquefaciens* strains were discriminated by variations in their partial *cheA* and *gyrA* sequences. Branching of the two clades was visible in the Neighbour-Joining (NJ) phylograms and was supported by bootstrap values of 76% and 100%, respectively. However, variations in selected marker gene sequences are not sufficient for discriminating taxonomic categories and for establishing novel subspecies. Therefore, we used several genomic methods, e. g. direct whole genome comparison, digital DNA-DNA hybridization, and microarray-based comparative genomic hybridization (M-CGH) as complementary approaches to justify that both ecovars represent two different subspecies. These methods will be described in more detail in course of this chapter. The known genome sequence of plant-associated FZB42 (Chen *et al*., 2007) and the novel whole genome sequences obtained from *B. amyloliquefaciens* type strain DSM7^{T} (Rueckert *et al.,* 2011) and of three Chinese plant –associated *B. amyloliquefaciens* strains, known for their potential to promote plant growth, were included in our analysis. The differences detected in our genome comparisons, especially deviations in the core genomes, changes in the variable portion of the genomes, differences in values obtained in DDH and MCGH – patterns were indicative for discriminating the members of the FZB42 subgroup (*B. amyloliquefaciens* subsp. *plantarum* subsp. nov.) and the strains related to the *B. amylolique-*

faciens type strain DSM7 (Borriss *et al*., 2010).

II. Sequencing techniques: Next Generation genome sequencing

A. Sequencing techniques

The key technology to enable taxonomic studies on the level of whole genomes respectively proteomes was the introduction of the so called next generation sequencing (NGS) technologies. Before the advent of these technologies, establishing a complete genome sequence using the classical Sanger sequencing approach required a huge amount of lab work to prepare the necessary clone libraries and a high amount of sequencing time, which prevented the widespread use of whole genome sequencing for taxonomic purposes.

1. Next generation sequencing

With the commercial introduction of two platforms for high-throughput sequencing, the determination of the whole genome sequences of several strains of a certain species or several species of a genus for taxonomical purposes alone has become feasible. Both, the Genome Sequencer (GS) by 454 Life Sciences (Branford, CT, U.S.A.) and the Genome Analyzer (GA) by Solexa (San Diego, CA, U.S.A.) get rid of the need of clone libraries, as both rely on PCR-based library preparation techniques. Initially, both systems were not really suitable for *de novo* sequencing as only 20 to 30 bases could be reliably determined. As of 2011, the obtainable length has increased to about 450 bases with the GS-FLX platform (with Titanium reagents) and 2x 150 bases with the GA-II*x* system, thus reaching a range reminiscent of the early automated Sanger sequencers. In contrast to the latest Sanger sequencers which could sequence only up to 384 samples in a single run, the NGS platforms provide millions (GS-FLX) to hundreds of millions (GA-II*x*) of sequences per run, driving the cost per assembled Mbase well below 1.000$.

While both systems allow *de novo* sequencing, several practical considerations have to be taken into account when deciding which one to use for a “whole” genome project, due to the strengths and weaknesses of the two systems, as discussed below.

a. Pyrosequencing (Genome Sequencer FLX, 454/Roche)

Being the first commercial NGS system in widespread use, the Genome Sequencer is based on the principle of pyrosequencing, first described by Ronaghi *et al.* (1996). Instead of using fluorescently labeled nucleotides or primers, the sequence read-out occurs via the conversion of pyrophosphate to ATP which is in turn converted to light by firefly luciferase. As with Sanger sequencing, it took almost a decade to create a viable commercial platform usable for whole genome sequencing. After a long series of optimizations and especially a high degree of miniaturization, the GS platform was introduced in 2005 (Margulies *et al.*, 2006).

Today, the GS-FLX platform with Titanium reagents allows to sequence approximately 500 Mbases in a single run taking about 5 hours. This allows the *de novo* sequencing of a (hypothetical) bacterial genome of 20 million base pairs length with a coverage of 25-fold which is usually sufficient to correctly assemble 95-99% of the sequence with good quality. In standard practice, the picotiter plate used by the GS-FLX is segmented to allow two, four or eight bacterial genomes to be sequenced in parallel without tagging. While cost constraints normally prohibit the acquisition of a GS-FLX (or a GA-II*x*) by individual laboratories, the service can easily be bought from specialized companies and institutions. In addition, Roche has recently launched the GS Junior which is suitable for “small scale” sequencing, i.e. a few dozen bacterial genomes per year.

When compared to the GA-II*x*, the main advantage of pyrosequencing with the GS-FLX system lies in the read length. While 450 bases compared to 2 times 150 bases does not appear to be much of a difference, one has to keep in mind that the subsequent sequence assembly will be interrupted by repetitive elements of approximately the size of the read length. Thus, the assembly of GS-FLX reads usually results in far fewer contigs than an assembly obtained from GA-II*x* reads. Another advantage, at least for researchers without a strong background in bioinformatics and without access to powerful compute clusters is the fact that the GS-FLX system comes with its own assembly software, including a graphical user interface (GUI). Together with the superior quality of the read data, this usually allows the assembly of a draft genome consisting of a few to a few hundred contigs within a few hours. This is usually sufficient for many genomic and taxonomic studies, e.g. to inventory up to 99% of the protein coding genes, to calculate the core and pan genomes, etc. (see Section III for details). For establishing the complete genome sequence, the GS-FLX *de novo* assembler allows to automatically recognize and process reads from long paired-end libraries of up to 20 kbp. Thereby, usually all unique contigs of a genome can be assembled into one or a few scaffolds (arrays of contigs with known order) from a single library and single nucleotide polymorphisms in repetitive elements can often be resolved without the need for additional Sanger sequencing. In addition, the gsAssembler software can provide output in ACE-format, which facilitates subsequent finishing approaches (see Section II.B).

On the downside, the per base pair cost for GS-FLX sequencing is (as of early 2011) approximately 100fold that of GA-II*x* sequencing, which usually prohibits large scale sequencing of dozens or hundreds of genomes for purely taxonomic purposes. Another drawback of this technology is the so called homopolymer problem. Pyrosequencing of long stretches (> 8 nucleotides) of identical bases results in often leads to an over- or underestimation of the correct number of the nucleotide in question. Therefore, GS-FLX sequencing genomes with an either a high G+C or a high A+T content often results in a higher number of frame-shifts due to the increased probability of longer homopolymer stretches. A related problem are sequence gaps that result from PCR biases introduced during library preparation. Extremes in G+C content (in either direction) result in an increased probability of hairpin-loop formation that can inhibit or completely prevent PCR amplification. As the emPCR necessary during GS-FLX library preparation cannot be optimized as rigorously as the normal PCR needed for GA-II*x* libraries, genomes with an extremely high G+C content tend to have many poorly covered or even uncovered regions after GS-FLX sequencing.

b. Sequencing by synthesis (Genome Analyzer IIx, Illumina/Solexa)

Based on the stepwise synthesis and subsequent detection of the incorporated nucleotide by fluorescence, the GA-II*x* platform allows sequencing of more than 200 million reads of up to 2x 150 nucleotides in a single run. While considerably slower than the GS-FLX (the above example would take about 14 days to complete) the amount of data obtained far exceeds the output of a GS-FLX run, ranging between 60 to 90 Gbases, an amount more than doubled with the recent introduction of the HiSeq2000 platform. This drives the price per Mbase of assembled sequence well below 100$, allowing for truly large scale studies. As mentioned in Section II.A.1.a, another advantage of GA-II*x* sequencing is the possibility to obtain less biased PCR libraries or completely forego PCR amplification (Kozarewa *et al.*, 2009), making it suitable for genomes with an extreme high G+C content.

The main drawback for researchers without access to bioinformatics support (people as well as compute capacity) is the lack of a company-supplied assembly software for GA-II*x* data. A number of open source programs is available, e.g. ABySS (Simpson *et al.*, 2009), MIRA (Chevreux *et al.*, 1999), and velvet (Zerbino and Birney, 2008), but installation and utilization of these programs usually requires experienced users. This also complicates genome ordering and finishing using long paired end libraries, as many programs cannot utilize that kind of data. As mentioned above, another drawback is the still rather short read length which might cause problems when assembling genomes with many repetitive regions. As the length of GA-II*x* reads has steadily increased during the last years, this disadvantage might be resolved in the near future.

2. *Sanger sequencing*

Although the principle technique was discovered more than three decades ago (Sanger and Coulson, 1975), the sequencing of complete genomes was not started until 1990 when development and availability of automated sequencing machines had reached a critical mass. Today, if Sanger sequencing is used at all for new genome projects, it is applied during the finishing and polishing phases when the gaps and low quality regions left by assembling NGS data are resolved. For these applications it remains a necessary and valuable technique for the next future.

B. Assembly, finishing and annotation

1. *Assembly and finishing*

The choice of the NGS technique(s) to use for a project depends heavily on the information the researcher wants to obtain for the taxonomic comparison(s). If one is only interested in studying the phylogeny based on core and/or pan genomes (see Section III.A.2), scaffolding of contigs and finishing is usually not necessary, so the NGS technique might be selected primarily on the basis of costs and the availability of bioinformatics resources (see Section II.A).

On the other hand, if more detailed studies, like genomic rearrangements or the distribution and movement of mobile genetic elements is of interest, use of the GS-FLX platform, gsAssembler and a 10 kbp long paired end library is strongly recommended, at least for the near future. This will usually result in one or two scaffolds per replicon and provide a file in ACE format for finishing and polishing. For the latter two steps, consed (Gordon *et al.*, 1998) is perhaps still the best finishing packages around. Developd as part of the Phred/Phrap package for the assembly and finishing of Sanger-based projects, consed has been updated to handle also GS-FLX and GA-II*x* data. Together with autofinish (Gordon *et al.*, 2001), the basic steps of finishing (primer selection, etc.) can easily performed based on the scaffolding data. The final task when attempting to completely close a genome is the creation/selection of suitable templates for Sanger sequencing. The straightforward approach relies on PCR to amplify the regions of interest, the indirect one consists of creating a large insert library and identifying of suitable templates by random sequencing or hybridization. Like selecting the best NGS technique for a project, choosing a technique for gap closure depends heavily on the circumstances like the organism to be finished, the number of gaps, and the genome size. PCR is most suitable when the number of gaps is small, the genome is large and the number of repetitive elements is low. As only the required templates are created, the amount of lab work is comparatively low. On the downside, all the usual problems of PCR like introduction of errors, formation of chimeric templates (in case of large repetitive regions), false priming, etc. can severely hinder this approach, especially when a huge number of gaps have to be addressed. These pitfalls are avoided by utilizing a large insert library, but creating and screening it is expensive in time and money. Therefore, it is usually only useful if a large number of gaps have to be addressed, especially if they are caused by complex repetitive elements.

2. Gene prediction and annotation

Once the genome sequence has been established, either complete or at least on the level of a suitable draft (median contig size of at least 10 kbp), a crucial step for phylogenetic studies based on whole genome data is the correct prediction of genes and, to a lesser extent, their correct annotation (i.e. prediction of function). A number of gene prediction tools have been developed over the years, which can roughly be divided in two classes: *ab initio* predictors and comparison-based predictors. The former rely on intrinsic signals in the DNA that allow the differentiation of protein coding and non-protein coding regions, the latter search for protein sequences that are similar to those of other organisms. Examples for *ab initio* predictors include the widely used GLIMMER (Delcher *et al.*, 1999) and Prodigal (Hyatt *et al.*, 2010). The drawback of the *ab initio* approach is that these programs tend towards overprediction (i.e many false-positives) and are often miss the correct start codon. Comparison-based predictors like CRITICA (Badger and Olsen, 1999) on the other hand can only find genes that are also present in other organisms and therefore tend to miss singletons. One way to combine the strengths of both approaches is by combining them in one tool, as demonstrated , e.g. for GISMO (Krause *et al.*, 2007), another way is to use several different tools and to weight and combine their output. The latter approach is, e.g., realized in the REGANOR web server (Linke *et al.* 2006).

A good resource for gene prediction and subsequent annotation is the GenDB genome annotation system (Meyer *et al.*, 2003). The software is open source and can be obtained either as a standalone system or utilized as a web-based service. In the latter case, a user management system allows to handle confidential (i.e. unpublished) genome data and no additional software or databases have to be installed respectively maintained by the user(s). GenDB can handle complete microbial genomes as well as draft sequences and offers a number of useful pipelines and features. For gene prediction, the REGANOR pipeline is used, which in turn utilizes GLIMMER and CRITICA to do the actual gene prediction, but can be expanded to use other predictors of the users choice. Once the coding regions in a genome have been identified, the Metanor pipeline can be used to predict and automatically annotate the functions of the encoded proteins. As the REGANOR pipeline, this pipeline uses the output of different tools like BLAST, hmmsearch, and SignalP run against different databases (e.g. nr, SwissProt, anf PFAM) to create an automated gene annotation with as much depth as possible. Once again, a user can specify additional tools and/or databases to be used, thus tailoring the results to the specific needs of the user. When the automated annotation is complete, the user can manually check and curate individual annotations, all of them are archived. This provides a reliable basis for further taxonomic analyses.

III. Comparative genome analysis

This chapter describes tools and techniques for the comparison of microbial genomes.

A. Genome comparison and phylogeny

1. Global alignment of genomes

A first genome based approach to gain insight into the evolutionary distance between two species is to inspect the synteny of the genome sequences are. Two popular tools dedicated for whole genome comparisons are MUMmer (Kurtz *et al.*, 2004) and the Artemis comparison tool ACT (Carver *et al.*, 2005):

- MUMMER – MUMmer (MUM = Maximum Unique Match) is an open source software package for the rapid alignment of large genomic sequences on DNA and amino acid level. It provides a wide range of tools and utilities for alignments, filter steps and result visualization that can be combined to analysis pipeline. A typical pipeline for the comparison of two complete genomes would consist of:
    - NUCmer – Basic nucleotide alignment of the two sequences
    - show-coords and show-aligns – Parsing of the alignment output of NUCmer
    - delta-filter – Filtering of the alignments by length, identity, consistency etc.
    - mummerplot – Plotting and graphical representation of alignment results

Figure 1 shows three synteny plots generated by this pipeline.

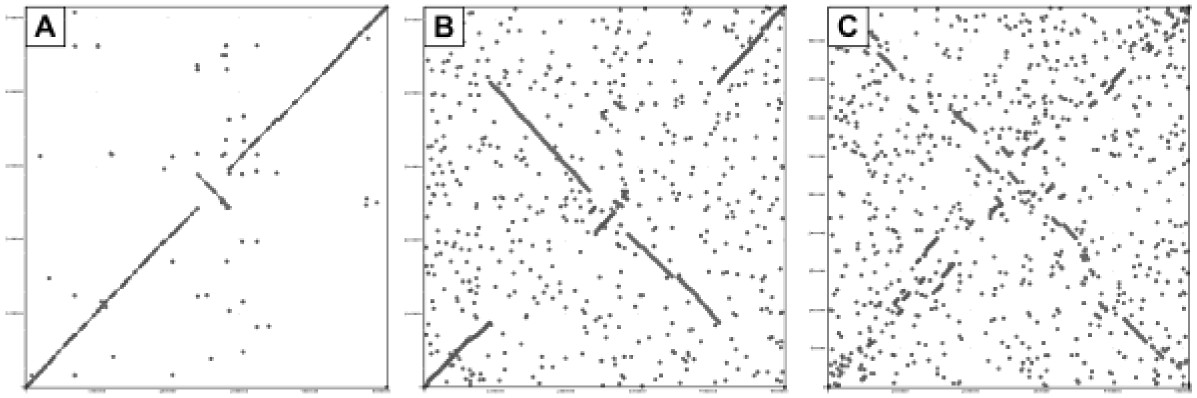


Figure 1. Synteny plots generated by MUMmer 3.07 comparing four *Xanthomonas campestris pv. campestris B100* with three strains from the same genus, *Xanthomonas campestris pv. Campestris 8004* (A), *Xanthomonas axonopodis pv. citri str. 306* (B), and *Xanthomonas oryzae pv. oryzae KACC10331* (C). A dramatic increase in genomic rearrangements can be observed, whilst synteny is decreased corresponding to increasing phylogenetic distance.

- ACT – The Artemis Comparison Tool ACT is a visualization software for pair wise genome alignments. It does not calculate alignments itself, but uses precalculated tabular output of the popular alignment software BLAST (Altschul *et al.,* 1997). Artemis allows the user to easily explore the synteny of the compared sequences and which genes are affected by genomic rearrangements.

2. Gene based genome comparison

Availability of completely sequenced and well annotated sequences for several members of a taxonomic group allows genotypic characterization of prokaryotes based on their gene content. Several terms have been defined in this field describing certain genomic subsets:

- **Core genome –** The term “core genome” denotes the set of genes that is shared by a group of analyzed organisms, usually all members of a certain taxonomic group (e.g. genus). This means that all genes of the core genome possess an orthologous gene in any other strain of the genome group.

- **Singleton genes –** The term “singleton gene” describes a gene that is unique within a group of analyzed organisms. This means that no orthologous genes can be identified in any other strain of the comparison set.
- **Pan genome** – The term “pan genome” specifies the set of all independent genes within a group of analyzed organisms. It comprises of the core genome, all singleton genes and all genes that can be identified in more than one, but not in all compared genomes. The pan genome describes the complete genetic potential of the analyzed taxonomic group.

Several tools and databases have been developed to identify groups of orthologs within a set of genomes and to calculate the genomic subsets mentioned above, for example the Comprehensive Microbial Resource (CMR), the Microbial Genome Database (MBGD), and EDGAR (Peterson *et al*, 2001; Uchiyama *et al.*, 2010; Blom *et al.*, 2009).

The CMR provides comparative tools for a database of 723 microbial genomes (64 of them draft genomes). E.g. the multi-genome homology comparison tool allows the user to calculate the number of homologous genes between up 15 selected comparison genomes. Special gene sets like the core genes or the singletons can be observed and exported in a tabular format. Another comparative tool included in the CMR is the genome homology graph, a dot plot showing the number of homologous genes between a selected reference genome and all genomes in the CMR database. The MBGD features comparative analyses for 1042 finished bacterial genomes. The genes of selected genomes can be clustered to homologous groups, resulting in a set of ortholog clusters. Additional analysis and visualization features are available for the clustered genes like multiple alignments or a comparison of the context of the genes on a genome map.

3. EDGAR

EDGAR, another resource for comparative genome analysis, is a dedicated approach for comparative and phylogenetic analysis of closely related genomes. EDGAR (Efficient Database framework for comparative Genome Analyses using BLAST score Ratios) provides several analysis and visualization features based on all-against-all BLAST comparisons of all genes of a set of analyzed genomes. EDGAR uses a generic orthology criterion adjusted to the set of compared genomes based on BLAST score ratios (Lerat *et al.*, 2003), a technique where every BLAST hit is weighted in relation to the maximum score. Based on this generic threshold EDGAR creates project specific databases storing the orthology information and serving as data source for subsequent analyses. EDGAR provides precalculated public databases for 95 bacterial genera with 846 genomes in total, but it is also possible to create private, access-controlled projects to analyse user-defined sets of genomes or unpublished data. Furthermore it is possible to create EDGAR projects directly from GenDB projects.

EDGAR features the calculation of the core genome, the pan genome and the singleton genes of all or subset of genomes included in a project. It is also possible to calculate specific gene sets by defining boolean operations on genomes. To visualize the distribution of shared and unique genes of compared genomes Venn diagrams of up to five genomes can be created (see Fig. 2). The comparative view provides a linear view of all orthologous genes in their genomic neighbourhood. To investigate large scale genomic events EDGAR provides an interface to create synteny plots (Fig. 1).

Furthermore EDGAR supports the differentiation between open and closed pan-genomes (Medini *et al.*, 2005) by predicting the number of singletons introduced by each genome with increasing genome number. For this purpose the number of singletons is calculated for each possible combination of genomes, subsequently a decay function is fitted to the averaged number of singletons for each quantity of genomes as described by Tettelin *et al.* (2005).

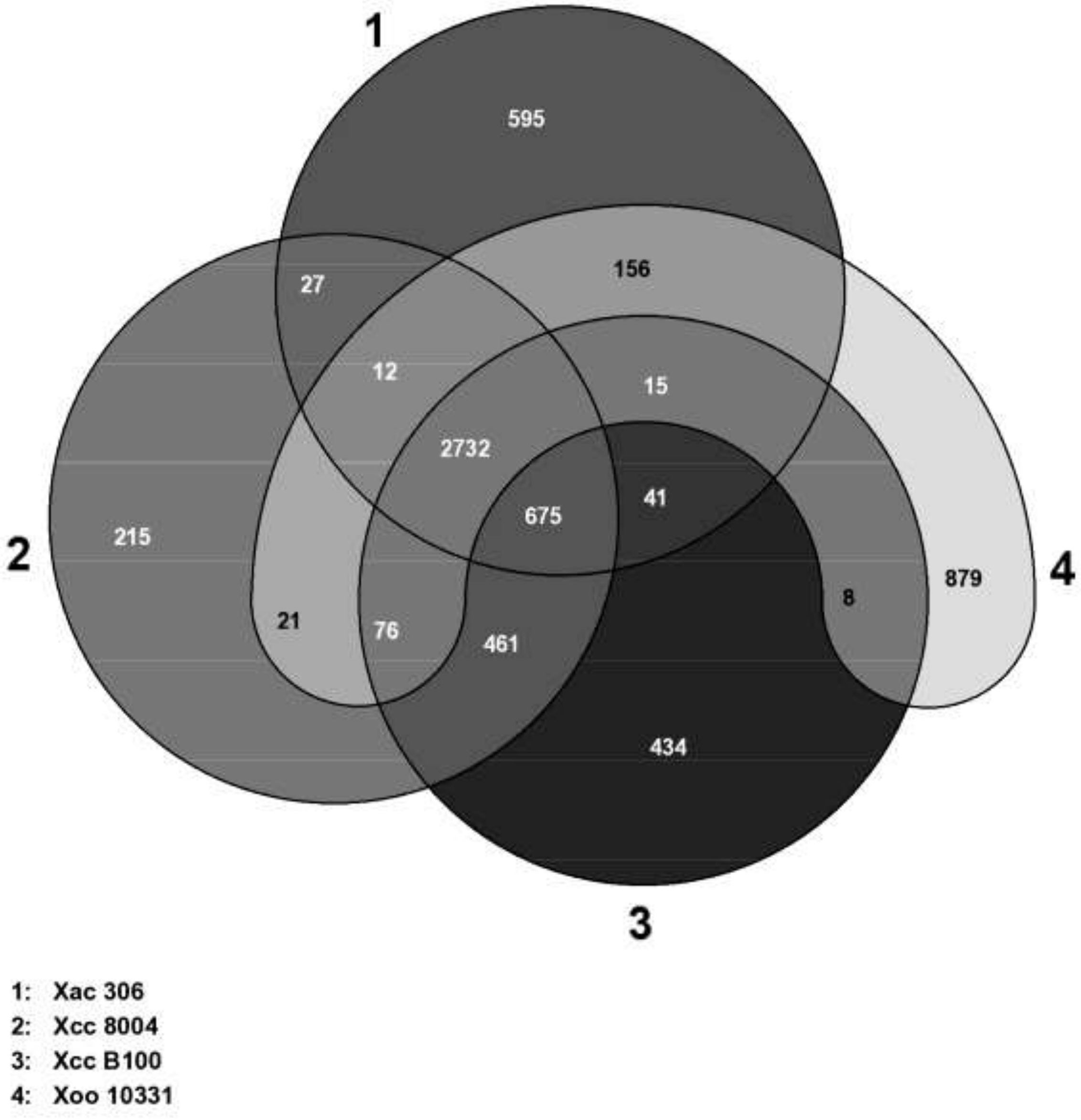


Figure 2: Venn diagram of the four *Xanthomonas* strains listed in Fig.1 showing the number of orthologous genes shared by the different strains.

4. *Phylogenetic trees*

Phylogenetic trees are branching diagrams illustrating the evolutionary relationships among species. Usually such trees are constructed based on sequence similarity between the highly conserved 16S rRNA genes or a set of housekeeping genes of several organisms. This limitation to a small set of input sequences can be problematic as the phylogeny of single genes does not necessarily reflect the phylogeny of the complete organisms. It is therefore highly desirable to use all genes of the core genome as input for the tree calculation, which increases dramatically its reliability (Gontcharov *et al*, 2004). EDGAR creates multiple alignments of all ortholog-sets of the core genome by using MUSCLE (Edgar, 2004), removes unaligned parts with GBLOCKS (Talavera, 2007), concatenates the multiple alignments of the single genes to one large alignment and finally creates a phylogenetic tree with the neighbor-joining implementation of the PHYLIP package (Felsenstein, 1995).

PHYLIP is a comprehensive collection of software tools that implement various algorithms for the creation of phylogenetic trees. Four of the most prominent algorithms are:

- **UPGMA -** Unweighted Pair Group Method with Arithmetic Mean: A simple clustering method, that assumes a constant rate of evolution (molecular clock hypothesis). Needs a distance matrix of the analyzed taxa that can be calculated from a multiple alignment.
- **Neighbor-joining (NJ):** Bottom-up clustering method that also needs a distance matrix. NJ is a heuristic approach that does not guarantee to find the perfect result, but under normal conditions has a very high probability to do so. It has a very good computational efficiency, making it well suited for large datasets.
- **Maximum parsimony (MP):** This method tries to create a phylogenetics that requires the least evolutionary change. It may suffers from long branch attraction, a problem that leads to incorrect trees in rapidly evolving lineages (Felsenstein, 1978).
- **Maximum-likelihood (ML):** ML uses a statistical approach to infer a phylogenetic tree. ML is well suited to the analysis of distantly related sequences, but is computationally expensive and thus not that well suited for larger input data.

While phylogenetic trees calculated from large sets of orthologous genes are quite reliable, trees generated from smaller samples may need some further confirmation. In such cases the use of an outgroup and further bootstrapping support can be helpful:

- **Outgroups:** When using distance matrix methods it is highly recommended to include at least one distantly related sequence to the analysis. This usage can be seen as a negative control, the outgroup should appear near the root of the tree and have a longer branch length than any other sequence.
- **Bootstrapping**: Bootstrapping is a resampling technique that is often used to increase the confidence that the inferred tree is correct. In a defined number of iterations (usually 100 – 1000) the multiple alignment that serves as input is permutated randomly and a phylogenetic tree is calculated. When the procedure is finished, a majority-rule consensus tree is constructed from the resulting trees of each bootstrap sample. The branches of the final tree are labeled with the number of times they were recovered during the procedure.

B. Electronic DNA-DNA hybridization (DDH)

The development of nucleic-hybridization methods, introduced into prokaryotic systematics from the 1960s onwards, has allowed the indirect comparison of gene sequences. DNA-DNA hybridization is applied, when strains share more than 97% 16S rRNA gene sequence identity. DDH-values not exceeding 70 % are considered as an indication that the tested organism belongs to a different species than the type strain(s) used as reference (Tindall *et al.*, 2010).

In recent years, genome based “*in silico*” alternatives to the cumbersome “wet lab” experimental DDH estimate were developed. There are several indexes that are obtained by comparing pairwise genomes that could be used in taxonomy. Noteworthy are the Average

Nucleotide Identity (ANI; Konstantinidis *et al*., 2006) and Maximal Unique Matches (MUM; Deloger *et al*., 2009) indexes as they have been hypothesized to be able to substitute for DDH. ANI has been demonstrated to correlate with DDH, where the range of ~95–96% similarity may reflect the current boundary of 70% DDH similarity (Goris *et al.,* 2007). A genome-to-genome distance comparison (GGDC), has recently been developed (Auch et al. 2010a and 2010b). The method based on whole genome data and allows also including unfinished draft genome sequences. We took advantage of this method to determine genomic distances of FZB42 and DSM7^T. The complete *B. subtilis* 168 genome and the draft genomes of three further plant associated strains related to *B. amyloliquefaciens,* YAU Y2, CAU B946, and NAU B3, were also included in that analysis. The results demonstrated that *B. subtilis* and *B. amyloliquefaciens* can be discriminated on species level by their digital DDH values which are much lower than 70%, whilst the DDH values between FZB42 and DSM7 were calculated as being around 77% (Table 1, Borriss *et al*., 2010). Values, ranging between 70- 80%, are considered as sufficient for discriminating subspecies. GGDC analysis of the three draft genomes yielded DDH values of 86-88 % with FZB42, but only 74-77% when compared with

Table 1: „*In silico*“ GGDC analysis of genomic *Bacillus* DNAs using program BLAT. Regression-based DDH estimates (in %) are indicated (according to Borriss *et al.*, 2010).

Query/reference	formula	DSM7	FZB42	*B.subt.*168
B. amyl. DSM7 length=3980199 bp FN597644	1[1] 2[2] 3[3]	97.71660 87.07480 >=100	80.3271741 64.4133336 77.6359098	37.1054714 14.9453016 30.2944702
B. amyl. FZB42 length=3918589 bp CP000560.1	1 2 3	80.3271741 64.4133336 77.6359098	97.71660 87.07480 >=100	38.2065793 14.9819160 31.1536318
B. amyl. YAU-Y2 length=4198660 bp	1 2 3	75.9969741 64.1295630 73.6625246	84.561258 79.119362 85.7222127	35.5791967 14.2115092 28.9901377
B. amyl. NAU-B3 length=4154898 bp	1 2 3	77.244344 64.0638979 74.7654545	86.316703 79.1309892 87.3710331	36.4790607 14.3259432 29.7046335
B. amyl. CAU-B946 length=3978635 bp	1 2 3	80.761889 63.8177953 77.8597081	88.0674375 77.181397 88.4282404	35.2256760 13.9677772 28.6752984
B. subt. 168 length=4215606 bp AL009126.3	1 2 3	37.1054714 14.9453016 30.2944702	38.2065793 14.9819160 31.1536318	97.71660 87.07480 >=100

[1] Formula: 1 (HSP length / total length)

[2] Formula: 2 (identities / HSP length)

[3] Formula: 3 (identities / total length)

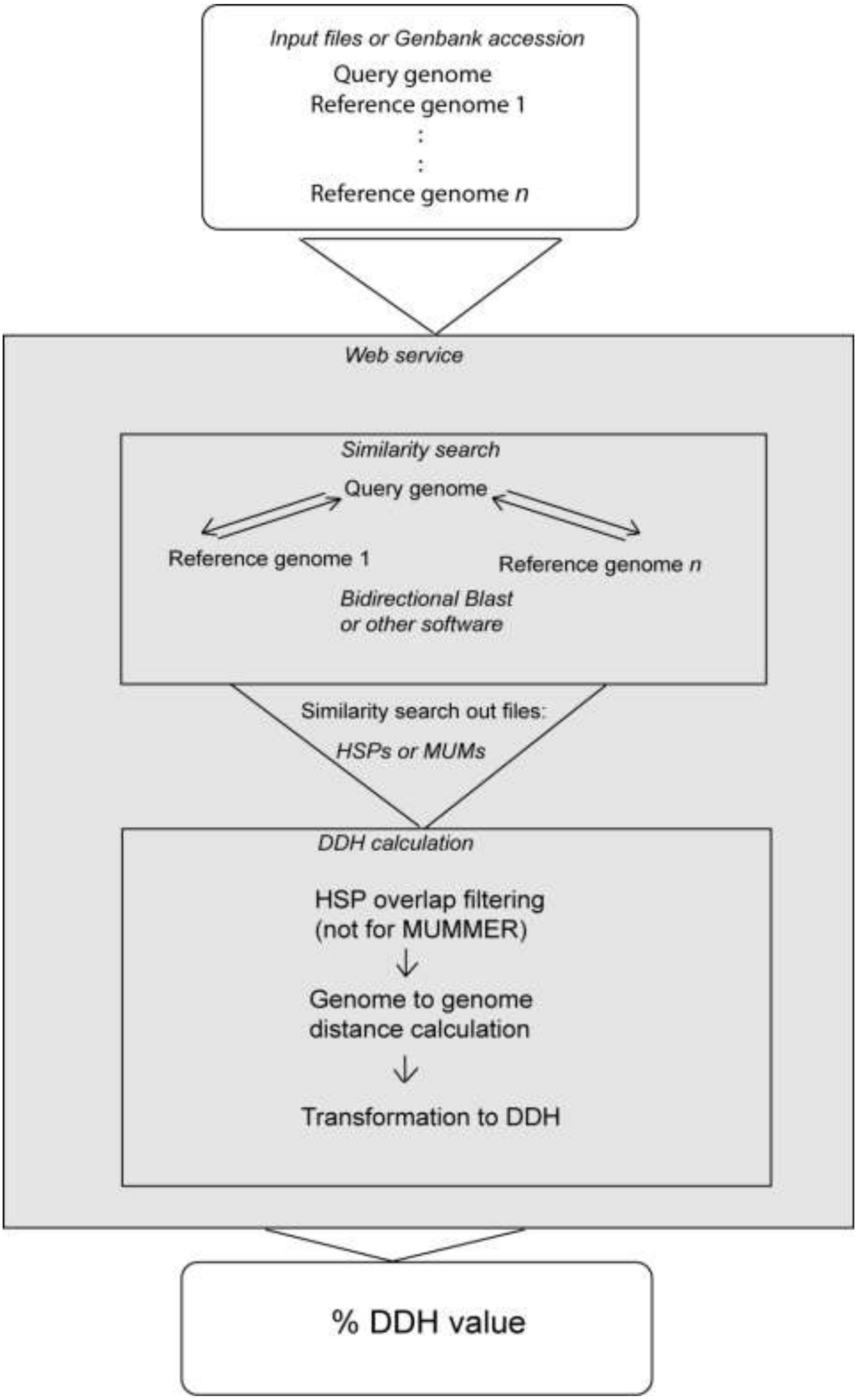


Figure 3. Flowchart outlining the steps required to calculate *in-silico* DDH values. Either Genbank accession numbers or FASTA files are uploaded on the server. The final values are received via e-mail (redrawn after Auch *et al.*, 2010a).

B. amyloliquefaciens DSM7^T strain, supporting closer taxonomic relatedness of the plant-associated *B. amyloliquefaciens* strains involved in our analysis (Borriss *et al,* 2010).

1. Genome-to-genome distance comparison (GGDC)

The main steps are: (1) determination of a set of HSPs or MUMs between two genomes, (2) the calculation of distances from these sets, and (3) the conversion of these distances in percent-wise similarities analogous to DDH (Fig. 3).

a. Requirements:

The GGDC web server (http://ggdc.gbdp.org) uses multi-FASTA files as input. One file per genome is expected, containing each chromosome or plasmid as a single FASTA entry.

A single query genome can be compared to several reference genomes; organism names can be entered separately. The user can choose between several similarity search tools. Presentation of the results is done via an e-mail to a user-specified address. The message also contains a brief explanation of the results (Auch *et al,* 2010a).

b. Similarity search

Similarities between query and reference genomes are determined by using well-known tools for nucleotide-based sequence similarity search. Currently, NCBI-BLAST, WU-BLAST (Altschul *et al,* 1990), BLAT (Kent, 2002) BLASTZ (Schwartz *et al*., 2003), and MUMmer (Kurtz *et al*., 2004) are available on the web server. High-scoring segment pairs, HSPs, or maximally unique matches, MUMs, are determined by performing similarity searches for each combination of query genome and reference genome. Due to the asymmetric nature of heuristic similarity search strategies, the search is performed twice, first using the reference genome as 'subject sequence' and the query genome as 'query sequence', and second, using the reference genome as 'query sequence' and the query genome as 'subject sequence'. The HSPs (or MUMs) are stored in condensed form using the CGVIZ format (Henz *et al.,* 2005), which comprises the start and stop coordinates of the matches together with statistical data (e-value, score, alignment length, and percentage identical characters for HSPs, alignment length for MUMs). The resulting data is sufficient for the distance calculation, while preserving storage space (Auch *et al*., 2010a).

c . Distance calculation

Distances between genomes are calculated using GBDP. When using NCBI-BLAST, WU-BLAST, BLAT, and BLASTZ, the gree-dy-with-trimming algorithm (Henz *et al*., 2005) is applied using distance functions. Distances for MUMmer are calculated using the coverage algorithm (Henz *et al*., 2005) with distance function. Considering error ratios and correlation with DDH, distance functions are recommend (Altschul *et al*., 1990). Filtering of HSPs having an e-value above 10^{-2} should be applied for BLAT, NCBI-BLAST and MUMmer prior to distance calculation, while it is not necessary for BLASTZ and WU-BLAST. A downstream filtering step has the advantage that it can easily be changed without the necessity to re-run the costly similarity search with adapted parameters. This enables one to reuse the data for further processing (Auch *et al*., 2010a).

d. Conversion to percent-wise similarities

The obtained distance values are converted into percent-wise similarities by using the corresponding values for intercept and slope. The percent-wise similarity can be used analogous to a DDH value. Values for intercept and slope are determined by applying the robust line fitting procedure as implemented in the R package (Version 2.6.2) to the dataset described in Auch *et al* (2010b).

$s(d)=md+c$

Additionally, the corresponding distance thre-shold as determined in (Auch *et al*, 2010b) can be used for species delimitation. Any distance value above the thre-shold can be regarded as indication that the two genomes analyzed represent two distinct species (Auch *et al*., 2010a).

C. Identification of horizontally transferred genomic islands

The genetic exchange was found to have occurred in different domains of life: Archaea, Bacteria, and Eukarya (Choi *et al*., 2007). Mobile genetic elements possess genes that contribute to bacterial speciation and adaptation to different niches (Dobrindt *et al.*, 2004). Collectively, the latter factors form part of a gene organization known as the *flexible gene pools*. The flexible gene pools are named according to the types of functions they encode, and are as follows:

- **Pathogenicity islands (PAI)** were first identified in uropathogenic *E. coli* strains as distinct chromosomal regions in possession of genes encoding virulence factors (Oelschlaeger *et al*., 2002). These factors enable bacteria to undergo several host-cell infection cycles, particularly, to adhere to host surfaces, attain protection against immune cells, and produce toxins. Virulence factors are disseminated by plasmids and bacteriophages, for they play the most crucial role in mobilizing virulent cassettes across species boundaries (Betley *et al*., 1985; Leplae *et al*., 2006; Lima-Mendez *et al.*, 2008a).
- **Symbiosis islands** share similar structural properties with pathogenicity islands. They both use similar mechanisms that influence the integration and host-bacterial interaction. Unlike pathogenicity islands, symbiosis islands are not associated with bacterial virulence. They encode new proteins and functions that establish mutual relationships between bacteria and multicellular organisms. For example, *Mesorhizobium* carry chromosomally integrated nitrogen fixation islands that benefit their plant hosts (Uchiumi *et al.*, 2004).

- **Antibiotic resistance islands** endow bacteria with multiple drug resistance. Bacteria can either develop the resistance by random mutations, transformation or transduction, but the most common way through which bacteria acquire drug resistance gene cassettes is conjugation. For example, most of the tetracycline resistance genes are identified in resistance plasmids, making horizontal transfer the likely method of their transfer (Hartman *et al.*, 2003; Pezzella *et al.*, 2004).
- **Catabolic genomic islands** possess genes that enable bacteria to degrade xenobiotic chemicals that are difficult to consume or even harmful to living organisms. Genes encoding relevant enzymes frequently have been found to be located within these islands (Butler *et al.*, 2007).

The transfer of genomic islands occurs through three mechanisms: transformation, conjugation and transduction. Upon transfer, these genetic elements get established into the recipient cell either as self replicating elements such as plasmids or by getting integrated into the chromosome (Dutta *et al*., 2002) either by homologous or illegitimate recombination techniques (Beiko *et al.*, 2005). The identification of genomic islands falls mainly on the basis of compositional features that distinguish them from native genes in the genome or they may be predicted by sequence similarity with previously identified genomic islands stored in databases.

1. Horizontal Gene Transfer Database (HGT-DB)

HGT-DB (http://genomes.urv.cat/HGT-DB/) is a composition-based web resource that provides pre-calculated averages and standard deviations for GC content, codon usage, relative synonymous codon usage and amino acid content of bacterial and archaeal complete genomes. It also provides lists of putative genomic islands, correspondence analyses of the codon usage and lists of extraneous genes in terms of their GC contents (Garcia-Vallve *et al.*, 2003). It uses a set of statistical approaches to determine the genes that deviate from the mean

GC and/or average codon usage of the genome. HGT-DB provides no tools for analysis of genomes submitted by users.

2. Pathogenicity Island Database (PAI-DB)

PAIDB (http://www.gem.re.kr/paidb/) contains the comprehensive information of all reported and potential PAI regions in prokaryotic genomes. In total 1040 PAI-like regions were identified in 237 bacterial genomes by the PAI Finder tool. PAI Finder accepts input sequences of predicted ORFs in multi FASTA format. The query is limited to 400 ORFs per run (approximately 350 kb). The PAI-DB resource may be used as follows:

- Predicted ORFs must be saved in a FASTA file and each sequence in the file has to be named strictly according to the PAI Finder format: ORF id, name, coordinates in the genome (left..right) and the strand (+/-) separated by vertical lines (|). For example:

  ```
  >3|name3617|3406225..3406300|+
  ATGCGGATAGCTCAGTCGGTAGAGCAGGGGATTGAAAATCCCCGTGTCCT
  TGGTT
  ```

- Query sequences of total length below 30 kbp may be pasted in text box on the web page, otherwise the sequences have to be stored locally and uploaded to the server.
- Click the button ‘Analyze’. The service returns lists of PAIs homologous to ones found in PAIDB for each ORF in the input.

As any other homology based prediction tool, PAI Finder has limitations: it may only identify PAI if at least one similar genomic island is already present in the PAIDB. However, PAIDB is regularly updated that thus improves its reliability. Other limitation is that the genome of interest has to be pre-annotated and a PAI may be overlooked if the annotation is not appropriate. Also preparation of the input file may be time consuming.

3. A Classification of Mobile Genetic Elements (ACLAME) project

Aclame (http://aclame.ulb.ac.be/) is a comprehensive web resource that aids with the classification and annotation of proteins encoded by mobile genomic elements (MGEs) (Leplae *et al.*, 2004). It has a collection of protein families obtained from bacteriophages and plasmids. The proteins were clustered into families according to functional parameters they have in common by using TRIBE-MCL, a graph theory based Markov clustering algorithm. Genomic islands were identified by using Prophinder tool (Lima-Mendez et al., 2008b). Prophinder was designed to detect prophages in bacterial genome sequences stored in GenBank formatted files (these files usually have extensions GB or GBK). Prophinder is available as an on-line tool and may be utilized as follows:

- Prepare a GenBank file of a bacterial genome of interest. GenBank files of many sequenced genomes are available for download from ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.
- Go to the Prophinder home page http://aclame.ulb.ac.be/perl/Aclame/Prophages/prophinder.cgi.
- Use the button 'Browse' to upload the genome file.
- Accuracy of the analysis may be adjusted by setting the scanning window size; minimum number of phage related CDS in prophages; minimal number of ACLAME hits per scanning window; and Blast e-value threshold. Prophinder is homology based approach. It predicts prophages by blasting the predicted proteins encompassed with the sliding window against ACLAME database of phage associated proteins.
- The prediction may be refined by secondary search after masking all obvious hits (set by default) and by looking for flanking repeats.
- Provide your e-mail for the server feedback and click the button 'Submit genome'.

The extreme mutability of phage related genes in prophages may make them undetectable by Blast search. Fragmentation of prophages due to genome rearrangements complicates the detection by Prophinder even greater.

4. IslandViewer

IslandViewer (www.pathogenomics.sfu.ca/islandviewer/) is a web-resource that incorporates precomputed genomic islands that were identified by the three prediction methods: IslandPick (Langille *et al.*, 2008), IslandPath (Hsiao *et al.*, 2003) and SIGI-HMM (Waack *et al.*, 2006). It provides a simple view of all genomic islands predictions for the latter methods through a single integrated interface.

To analyze the sequence of a newly sequenced bacterial chromosome, first write it to the file in GenBank or EMBL format. Then follow these steps:

- On the project web-site click 'Genome upload'.
- Choose the corresponding sequence file format and click the button 'Browse' to locate the file on the computer. Optionally the genome name may be entered to facilitate the navigation through the resulted graphs if multiple genomes are going to be analyzed.
- Click the button 'Upload'. When sequence upload is complete, enter the e-mail address and click 'Submit'. In a while you will be notified when the analysis is finished.
- Inspect your mail box. Eventually you will get a message with a hyperlink that will bring you to the result of the analysis. The locations of identified genomic islands predicted by IslandPick, SIGI-HMM and IslandPath will be depicted by green, orange and blue boxes, respectively. A high resolution graphical file and exact coordinates in an Excel file are available for download.

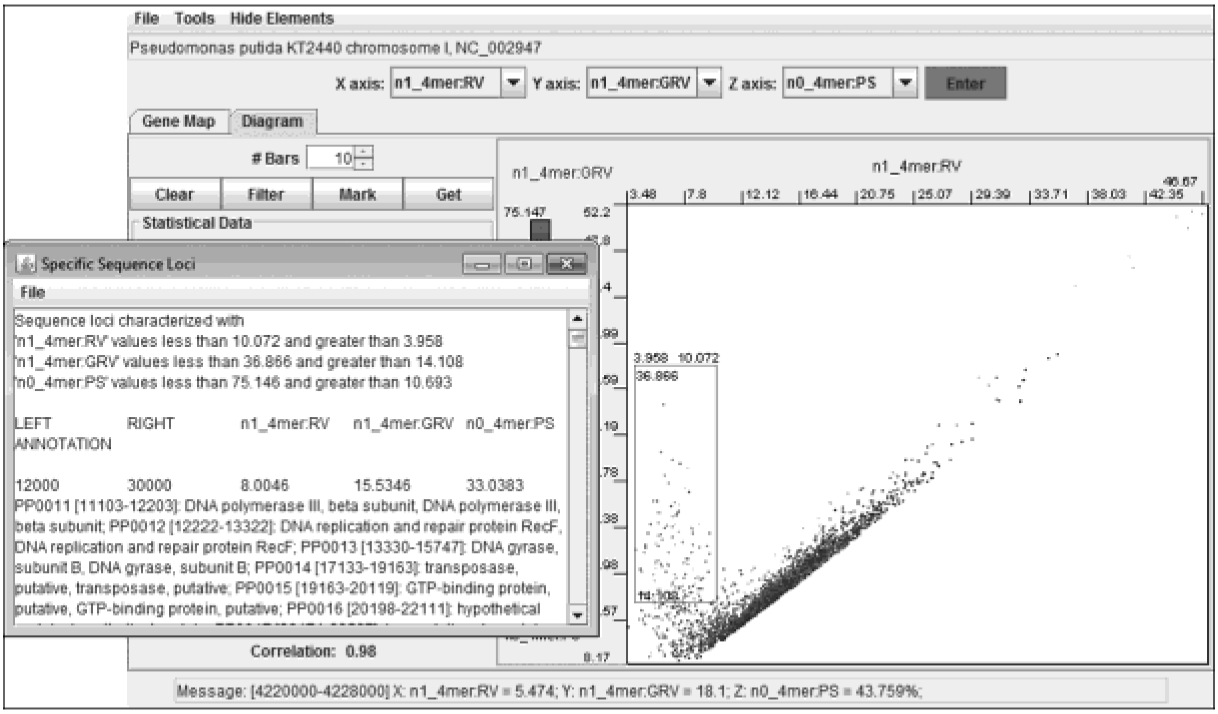
IslandViewer is superb in genomic island prediction by combining of three alternative approaches based on genome comparison (IslandPick), codon usage comparison using a Hidden Markov Model algorithm (SIGI-HMM) and DNA composition comparison algorithm (IslandPath).

5. SeqWord Genome Browser and Gene Island Sniffer

SeqWord Genome Browser (SWGB) was developed to visualize the natural compositional polymorphism of DNA sequences and to identify divergent genomic regions including horizontally transferred genomic islands (Ganesan *et al.*, 2008). The approach is based on the analysis of biased distribution of tetranucleotides in bacterial genomes. Several statistical parameters, – distances between local oligonucleotide usage (OU) patterns calculated for sliding windows and the global pattern of the whole genome, OU variance and pattern skew defined by Reva and Tümmler (2004, 2005), – are superimposed by the program to distinguish between mobile genomic islands and other elements characterized by an alternative OU (clusters of genes encoding ribosomal RNA and proteins, tandem multiple repeats, and so on). SWGB allows visual identification of genomic islands by browsing bacterial chromosomes, grouping genomic fragments by their compositional properties and a simultaneous referring to the genetic context. The SWGB resource may be used as follows:

- On the SWGB web-page (www.bi.up.ac.za/SeqWord/mhhapplet.php) select one bacterial chromosome, plasmid or phage in the list and click ‘Display in the Applet’;
- Click on the tab ‘Diagram’ and choose the parameters n1_4mer:RV, ni_4mer:GRV and n0_mer:PS for the axes X, Y and Z, respectively (More about these parameters and abbreviations see in Ganesan *et al.*, 2008). Click ‘Enter’. A diagram of distribution of 8 kbp long genomic fragments will appear on the plot. Using the mouse, draw a box around the group of dots on the plot (Fig. 4) and click ‘Get’. The

program will return a list of genomic loci and annotations represented by the outlined dots on the plot, which correspond to the horizontally transferred genetic elements.



Pseudomonas putida KT2440 chromosome I, NC_002947

X axis: n1_4mer:RV Y axis: n1_4mer:GRV Z axis: n0_4mer:PS

Bars 10

```
Sequence loci characterized with
'n1_4mer:RV' values less than 10.072 and greater than 3.958
'n1_4mer:GRV' values less than 36.866 and greater than 14.108
'n0_4mer:PS' values less than 75.146 and greater than 10.693

LEFT          RIGHT          n1_4mer:RV     n1_4mer:GRV   n0_4mer:PS
ANNOTATION

12000         30000          8.0046         15.5346       33.0383
PP0011 [11103-12203]: DNA polymerase III, beta subunit, DNA polymerase III,
beta subunit; PP0012 [12222-13322]: DNA replication and repair protein RecF,
DNA replication and repair protein RecF; PP0013 [13330-15747]: DNA gyrase,
subunit B, DNA gyrase, subunit B; PP0014 [17133-19163]: transposase,
putative, transposase, putative; PP0015 [19163-20119]: GTP-binding protein,
putative, GTP-binding protein, putative; PP0016 [20198-22111]: hypothetical
```

Correlation: 0.98

Message: [4220000-4228000] X: n1_4mer:RV = 5.474; Y: n1_4mer:GRV = 18.1; Z: n0_4mer:PS = 43.759%;

Figure 4. Identification of genomic islands in *Pseudomonas putida* KT2440 using SWGB.

- Double-click a dot on the plot to refer to genetic content of the corresponding region on the 'Gene Map' diagram.

Sequences stored in FASTA or GenBank files may be analyzed by SWGB locally.

- The command line Python program OligoWords is available for download from www.bi.up.ac.za/SeqWord/downloads.htm in several packages containing precompiled executable files. Input sequence files have to be copied to the 'OligoWords1.2.1\input\' folder.
- The command prompt window provides several parameters set by default which may be changed by the user (refer to the readme file). Type <Y>+<Enter> to perform the analysis.

- The program will analyze recursively all input files with the extensions FST, FAS, FNA and GBK, and store the results to the folder ‘output’ in text files with the extension OUT.
- Use the Java applet on the SWGB web-page to view these files using File->Open menu command.

SWGB allows composition based identification of genomic islands in annotated genomic sequences stored in GenBank files and in raw DNA sequences in FASTA format. The lengths of the sequences to be analyzed have to be above 20 kbp. SWGB is not able to identify inserts of foreign DNA shorter that the half of the sliding window size. One common problem for all sliding window based approaches is that the resulted prediction may depend on the starting point of the analysis. To improve and automate the prediction of genomic islands a SeqWord Gene Island Sniffer (SWGIS) utility was developed (www.bi.up.ac.za/SeqWord/sniffer/index.html). This command prompt program uses a shorter sliding window in the areas where an insertion is suspected. A collection of genomic islands identified in bacterial genomes by SW Sniffer is present in GEI-DB (http://anjie.bi.up.ac.za/geidb/geidb-home.php).

SWGB and Sniffer cannot identify genomic islands if they share similar OU distribution with the host chromosome. For example, symbiotic islands of *Rhizobium* are not detectable by SWGB. Clusters of genes for 16S rRNA are often falsely predicted as genomic islands. To identify all genomic islands in a chromosome the best way is to combine the results of different prediction methods, as suggested by Langille *et al.*, (2009). In Fig. 5 the results of predictions of horizontally transferred genetic elements in a newly sequenced genome of *Acidovorax avenae* ATCC 19860 by IslandPick, SIGI-HMM, IslandPath and SWGIS are superimposed on the chromosomal map.

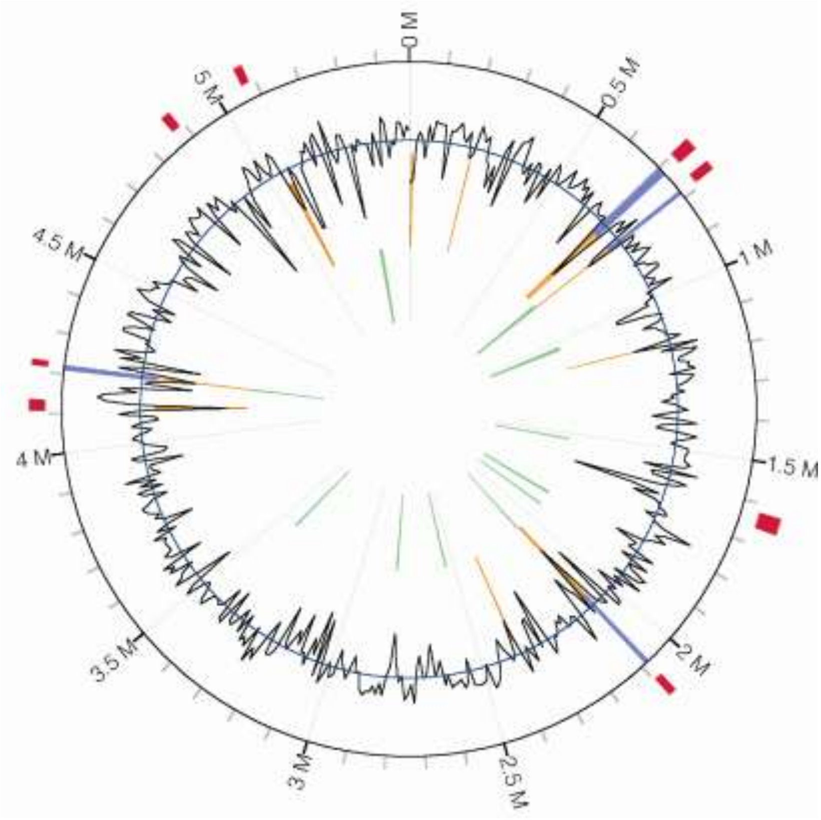


Figure 5. Identification of genomic islands in *Acidovorax avenae* ATCC 19860 by different prediction tools: SWGIS (red), IslandPath (blue), SIGI-HMM (orange) and IslandPick (green boxes). Black line histogram shows GC% variations.

IV. Microarray-based comparative genomic hybridization (M-CGH)

M-CGH is a powerful method for rapidly identifying regions of genomic diversity among closely related organisms in absence of complete genome data sets. The method can be applied in investigating a group of closely related strains given that a microarray prepared from a complete reference genome is available. The advantage is that no further whole genome sequencing of every strain under investigation is necessary. The disadvantage is that no genes unique in a single strain can be identified. The technique allows one to predict gene absence (or divergence) versus gene presence by measuring the relative hybridization efficiencies of two differentially Cy-labeled pools of genomic DNA taken from two strains. The method has been previously applied to discriminate different members of the *B. subtilis* clade (Earl *et al.,* 2007). We used *Bacillus amyloliquefaciens* FZB42 specific oligonucleotide microarray as reference to identify genes that are absent or divergent in the strains under investigation. As controls, FZB42 – FZB42, and FZB42 – CH40 (FZB42Δ*nrs*::*cm* and Δ*dhb*::*em*) hybridisations were performed. As expected, the self-self control experiments yielded no genes with a $\log_2$ fluorescence ratio greater than 1. The results of the FZB42-CH40 hybridization did, however, reveal a potential limitation of the array; the values for only 8 of the 11 genes known to be deleted in this strain were above the cutoff ratio for gene absence or divergence. This may have been a consequence of cross-hybridization between gene spots, because the missing genes *nrsA*, *nrsC* and *nrsF* contain peptide synthetase modules which exist quite frequently in the genome of FZB42. Strains FZB42, its derivative CH40 and FZB24 were nearly identical according to the heat map composite image (Borriss *et al.,* 2010). The other plant-associated *B. amyloliquefaciens* strains formed a cluster with only limited diversity in comparison to FZB42. DSM7^T, together with the two other non-plant-associated *B. amyloliquefaciens* strains, ATCC15841 and S23, were more diverse and formed a separate cluster. Interestingly, the phylogeny obtained when the degree and pattern of gene variation measured by the arrays was used as a marker of relatedness was in almost perfect concordance with the phylogeny obtained when *gyrA* and *cheA* was used as markers of relatedness. In both cases the cluster of *B. amyloliquefaciens* FZB42 related strains were discriminated from the cluster of DSM7^T related strains. As shown for *B. subtilis* (Earl *et al.,* 2007), M-CGH may also prove to be reliable phylogenetic tool for subtyping strains of *B. amyloliquefaciens* (Borriss *et al*., 2010).

A. Microarray-based comparative genomic hybridization – Procedure

1. *Hybridisation and microarray scanning*

Microarray slides were printed by the Center for Biotechnology, University of Bielefeld. *Bacillus amyloliquefaciens* FZB42 specific oligonucleotide microarray was applied to identify genes that are absent or divergent in the test strains. Each array was spotted with 3,931 50- to 70-mer oligonucleotides: 3,816 spots represented the FZB42's predicted gene set, 238 the hypothetical smRNAs (Chen et al. 2007). All oligonucleotides were spotted four times in a microarray plate. The comparative hybridizations were repeated for each test strain two or three times and included at least one hybridization where the labelling regimen was switched to rule out potential bias introduced by inherent differences in Cy dye incorporation. Five micrograms of purified, *Msp*I- and *Taq*I-digested genomic DNA was labelled with either Cy3- or Cy5-NHS ester as described by Giuntini *et al.*, 2005. Unincorporated fluorescent nucleotides were removed by using Microcon 30 filter columns (Millipore, Milano, Italy). The appropriate Cy5 and Cy3 labelled probes were combined and mixed with 30 µl Cot-1 DNA (1 mg/ml), 20 µl Yeast t-RNA (5 mg/ml), 450 µl TE to concentrate the samples until about 40 µl using Microcon 30 filter columns (Millipore,Milano, Italy). To each combined sample 8.5 µl of 20 × SSC and 0.74 µl of 10% SDS were added. The sample was denatured to 100°C for 1.5 min, and then incubated for 37°C for 30 min. The hybridisation probe was added to the microarray under a coverslip, and hybridisation was performed at 65°C for 16 h. Slides were washed at 60°C with 2 × SSC for 5 min and then at 60°C with 0.2 × SSC containing 0.1% SDS for 5 min and finally at room temperature with 0.2 × SSC for 2 min. The last step was conducted twice. The slides were immediately dried and scanned for fluorescence intensity by using a GenePix 4000B microarray scanner (Axon Instruments, Union City, CA), and the results were recorded in 16-bit multiimage TIFF files.

For each sample a total of four slides were hybridized (after dye swapping of the two different restriction enzyme DNA preparations); considering that one slide carries three replicas of each ORF, any sample was hybridized twelve times at each ORF.

2. Normalisation and significant hybridisation differences

Following hybridization and scanning, data analysis was done by applying the ImaGene 6.0 software (Biodiscovery Inc., Los Angeles, CA) for acquisition of the mean signal and mean local background intensity for each spot of the microarray and the EMMA 2.2 software for normalisation and *t*-statistics (Dondrup *et al*., 2003, 2009). A gene was considered to have a statistically significant difference in hybridisation if the $\log_2$-ratio of the intensities (M value) was ≥1 or ≤−1 and the mean intensity (A value; $A_i = \log_2(R_i G_i)^{0.5}$) was ≥7 and in two of the three repeats the P_{adjusted} value was ≤0.1. In this study a positive $\log_2$-ratio of the intensities (M value) indicated that the respective gene is missing in the genome of the tested strain.

3. EMMA and ArrayLIMS, useful platforms for microarray data processing

As a high throughput technique, microarray experiments produce large data sets, consisting of measured data, laboratory protocols, and experimental settings. CeBiTEC from University Bielefeld has implemented the open source platform EMMA http://www.cebitec.uni-bielefeld.de/groups/brf/software/emma_info/ to store and analyze these data. EMMA gives access to all the transcriptomics data sets stored in the ArrayLIMS and provides automated pipelines for data processing, allowing an automated or manual analysis of expression profiles. In addition to routine data analysis algorithms, the system can be integrated with other components that contain additional data sources (e.g., genome annotation systems). In the design of the microarray experiments, special care must be taken in projects within the same network, to ensure comparability of these data and compliance to new and arising international standards. This system also provides automated tools to perform data normalizations, tests for the identification of statistically significant up or down-regulated genes, clustering algorithms and, in the long run, support for time-course analyses.

ArrayLIMS is a Microarray Laboratory Information Management system has been designed in order to streamline data acquisition and reporting processes. It provides a permanent and consistent storage of the microarray experiment data as well as a fast information retrieval, making the data rapidly available. The stored data is standardized, consisting of the hybridization steps (e.g. RNA production), production of the hybridization targets or the hybridization itself. It is also possible to store images of the hybridized and scanned slides as well as the corresponding data files.

V. Concluding remarks

The drop in the price of sequencing whole genomes, together with the technical advances that have been made suggest that routine sequencing of prokaryote genomes is realistic from now on (Tindall et al., 2010). A key issue that remains is the reliable annotation of all genes in a genome since identifying gene homologies (preferably orthologues) is of central importance in taxonomy. In principle there are three basic approaches: (1) genome indexes, increasingly used as an "*in silico*" alternative to the experimental DDH, (2) gene content, its successful application depends on the number of genome sequences available for this analysis, and (3) multiple aligned (gene) sequence datasets (3). In this review we have presented several methods, we have found practicable for the non-experienced scientist with background in microbial taxonomy, for estimating those genomic parameters. We are sure that further development in the field will facilitate use of genomics as an essential part of prokaryote taxonomy.

Acknowledgement

Financial support in frame of the competence network Genome Research on Bacteria (GenoMikPlus) and the Chinese-German collaboration program by the German Ministry for Education and Research is gratefully acknowledged.

References:

- Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
- Altschul, S.F., Madden, T.L., Schäffer, 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. *Nucleic Acids Research* **25**, 3389-33402
- Auch, A. F., von Jan, M., Klenk, H.-P. and Göker, M. (2010a). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci*. **2**, 142-148.
- Auch, A. F., Klenk, H.-P. and Göker, M. (2010b). Standard operation procedure for calculating genome-to-genome distances based on high scoring segment pairs. *Stand. Genomic Sci.* **2**, 142-148.
- Badger, J.H. and Olsen, G.J. (1999). CRITICA: Coding Region Identification Tool Invoking Comparative Analysis. *Mol. Biol. Evol.* **16**, 512-524.
- Becker, A., Berges, H., Krol, E., Bruand, C., Rüberg, S., Capela, D., Lauber, E., Meilhoc, E., Ampe, F., de Bruijn, F. J., Fourment, J., Francez-Charlot, A., Kahn, D., Küster, H., Liebe, C., Pühler, A., Weidner, S. and Batut, J. (2004). Global changes in gene expression in *Sinorhizobium meliloti* 1021 under microoxic and symbiotic conditions. *Mol Plant Microbe Interact* **17**, 292-303

- Beiko, R. G., Harlow, T. J. and Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 14332-14337.
- Betley, M. J. and Mekalanos, J. J. (1985). Staphylococcal enterotoxin A is encoded by phage. *Science* **229**, 185-187.
- Blom, J., Albaum, S.P., Doppmeier, D., Pühler, A., Vorhölter, F.J., Zakrzewski, M. and Goesmann,A. (2009) EDGAR: A software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* **10**, 154
- Borriss, R., Chen, XH, Rueckert, C., Blom, J., Becker, A., Baumgarth, B., Fan, B., Pukall, R., Schumann, P., Sproer, C., Junge, H., Vater, J., Pühler, A. and Klenk, H.-P. (2010) Relationship of *Bacillus amyloliquefaciens* clades associated with strains DSM 7^T and *Bacillus amyloliquefaciens* subsp. *plantarum* subsp. nov. based on their discriminating complete genome sequences. *Int. J. Syst. Evol. Microbiol.* 2010 Sep 3. [Epub ahead of print]
- Butler, J. E., He, Q., Nevin, K. P., He, Z., Zhou and J. Lovley, D. R. (2007). Genomic and microarray analysis of aromatics degradation in *Geobacter metallireducens* and comparison to a *Geobacter* isolate from a contaminated field site. *BMC Genomics* **8**, 180.
- Chen, X.H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Heinemeyer, I., Morgenstern, B., Voss, B., Hess, W.R., Reva, O., Junge, H., Voigt, B., Jungblut, P.R., Vater, J., Süssmuth, R., Liesegang, H., Strittmatter, A., Gottschalk, G. and Borriss, R. (2007). Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nat. Biotechnol.* **25**, 1007-1014

- Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol.: Proc. German Conference on Bioinformatics GCB'99 GCB*: *45–56.*
- Choi, I.-G., and Kim, S.-H. (2007). Global extent of horizontal gene transfer. *Proc. Natl. Acad. Sci. USA* **104**, 4489-4494.
- Carver, T.J., Rutherford, K.M.,Berriman, M.,Rajandream, M.A.,Barrell, B.G. and Parkhill, J. (2005). ACT: the Artemis comparsion tool. *Bioinformatics* e21 (16), 3422-3423
- Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636-4641
- Deloger, M., El Karoui, M. and Petit, M.-A. (2009). A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J. Bacteriol.* **191**, 91–99.
- Dobrindt, U., Hochhut, B., Hentschel, U. and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol*. **2**, 414-424.
- Dondrup, M., Goesmann, A., Bartels, D., Kalinowski, J., Krause, L., Linke, B., Rupp, O., Sczyrba, A., Pühler, A., and Meyer, F. (2003). EMMA: a platform for consistent storage and efficient analysis of microarray data. *J. Biotechnol.* **106**, 135–146
- Dondrup M., Hüser A.T., Mertens D. and Goesmann A. (2009) An evaluation frame work for statistical tests on microarray data. *J Biotechnol*. **140**, 18-26
- Dutta, C. and Pan, A. (2002). Horizontal gene transfer and bacterial diversity. *J. Biosci*. **27**, 27-33.
- Earl, A. M., Losick, R. and Kolter, R. (2007). *Bacillus subtilis* genome diversity *J. Bacteriol.* **189**, 1163–1170.

- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1972
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat.l Acad. Sci. USA* **95**, 14863-14868
- Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology* 27(4), 401
- Felsenstein, J. (1995) PHYLIP (Phylogeny Inference Package), version 3.57 c. *Seattle: University of Washington*
- Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tummler, B. 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.
- Garcia-Vallve, S., Guzman, E., Montero, M. A. and Romeu, A. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res*. **31**, 187-189.
- Gatson, J. W., Benz, B. F., Chandrasekaran, C., Satomi, M., Venkateswaran, K. and Hart, M. E. (2006). *Bacillus tequilensis* sp. nov., isolated from 2000-year-old Mexican shaft-tomb, is closely related to *Bacillus subtilis*. *Int J Syst Evol Microbiol* **56**, 1475-1484
- Giuntini, E., Mengoni, A., De Filippo, C., Cavalieri, D., Aubin-Horth, N., Landry, C. R., Becker, A. and Bazzicalupo, M. (2005). Large-scale genetic variation of the symbiosis-required megaplasmid pSymA revealed by comparative genomic analysis of Sinorhizobium meliloti natural strains. *BMC Genomics* **6**,158

- Gontcharov, A.A., Marin, B. and Melkonian, M. (2004) Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the Zygnematophyceae (Streptophyta). *Mol. Biol. Evol.* **21**, 612
- Gordon, D., Abajian, C. and Green, P. (1998). Consed: A Graphical Tool for Sequence Finishing. *Genome Res.* **8**, 195-202
- Gordon, D., Desmarais, C. and Green, P. (2001). Automated Finishing with Autofinish. *Genome Res.* **11**, 614-625.
- Hartman, A. B., Essiet, I. I., Isenbarger, D. W. and Lindler, L. E. (2003). Epidemiology of tetracycline resistance determinants in *Shigella* spp. and enteroinvasive *Escherichia coli*: characterization and dissemination of *tet(A)*-1. *J. Clin. Microbiol.* **41**, 1023-1032.
- Henz S.R., Huson D.H., Auch A.F., Nieselt-Struwe K. and Schuster S.C. 2005. Whole-genome prokaryotic phylogeny. *Bioinformatics* **21**, 2329-2335.
- Hsiao, W., Wan, I., Jones, S. J. and Brinkman, F. S. L. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**, 418-420.

- Hsiao, W., Wan, I., Jones, S. J. and Brinkman, F. S. L. (2003). IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**, 418-420.

- Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.
- Kent WJ. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.* **12**, 656-664.
- Klappenbach, J. A., Coenye, T., Goris, J., Konstantinides, K. T.,Vandamme, P. and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91.

- Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philos Trans R Soc. Lond B Biol Sci* 361, 1929-1940
- Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*. **6**, 291-295
- Krause, L., McHardy, A.C., Nattkemper, T.W., Pühler, A., Stoye, J. and Meyer, F. (2007). GISMO--gene identification using a support vector machine for ORF classification. Nucleic Acids Res. **35**, 540-549.
- Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C. and Salzberg S.L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.
- Langille, M. G. and Brinkman, F. S. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**, 664-665.
- Lerat, E., Daubin, V. and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol* **1**, E19
- Logan, N. A., Berge, O., Bishop, A. H., Busse, H.-J., de Vos, P., Fritze, D., and 6 other authors (2009). Proposed minimal standards for describing new taxa of aerobic, endospore-forming bacteria. *Int. J. Syst. Evol. Microbiol.* **59**, 2114-2121
- Leplae, R., Hebrant, A., Wodak, S. J. and Toussaint, A. (2004). ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res*. **32**, D45-D49.
- Leplae, R., Lima-Mendez, G. and Toussaint, A. (2006). A first global analysis of plasmid encoded proteins in the ACLAME database. *FEMS Microbiol. Rev*. **30**, 980-994.

- Lima-Mendez, G., Helden, J. V., Toussaint, A. and Leplae, R. (2008a). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol*. **25**, 762-777.
- Lima-Mendez, G., Helden, J. V., Toussaint, A. and Leplae, R. (2008b). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24**, 863-865.
- Linke, B., McHardy, A.C., Neuweger, H., Krause, L. and Meyer, F. (2006). REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes. Appl Bioinformatics **5**, 193-198.
- Mantri, Y. and Williams, K. P. (2004). Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res*. **32**, D55-D58.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Goodwin, B.C., He, W., Helgesen, S., Ho, C.He, Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.Fredrik, Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. (2006). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature* **437**, 376-380.
- Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Devel.* **15**, 589-594

- Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. and Pühler, A. (2003). GenDB--an open source genome annotation system for prokaryote genomes. N*ucleic Acids Res.* **31**, 2187-2195.
- Oelschlaeger, T. A., Dobrindt, U. and Hacker, J. (2002). Pathogenicity islands of uropathogenic *E. coli* and the evolution of virulence. *Int. J. Antimicrob. Agents* **19**, 517-521.
- Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. and White, O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Research* **29**, 123-125
- Pezzella, C., Ricci, A., DiGiannatale, E., Luzzi, I. and Carattoli, A. (2004). Tetracycline and streptomycin resistance genes, transposons, and plasmids in *Salmonella enterica* isolates from animals in Italy. *Antimicrob. Agents Chemother*. **48**, 903-908.
- Reva, O. N. and Tummler, 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 Tummler, B. (2005). Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.
- Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**,84-89.
- Rooney, A. P., Price, N. P., Ehrhardt, C., Swezey, J. L. and Bannan, J. D. (2009). Phylogeny and molecular taxonomy of the *Bacillus subtilis* species complex and description of *Bacillus subtilis* subsp. *inaquosorum* subsp. nov. *Int. J. Syst. Evol. Microbiol.* **59**, 2420-2436

- Rueckert, C., Blom, J., Chen, X.H., Reva, O. and Borriss, R. (2011). Genome sequence of *B. amyloliquefaciens* type strain DSM7^T reveals differences to plant-associated *B. amyloliquefaciens* FZB42. *J. Biotechnol*. 2011 Jan 22. [Epub ahead of print]
- Sanger, F. and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448
- Schwartz S., Kent W.J., Smit A., Zhang Z., Baertsch R., Hardison R.C., Haussler D. and Miller W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107.
- Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117-1123.
- Talavera,G. and Castresana J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56,** 564-577
- Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.D., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L.,Wessels, M.R., Rappuoli, R. and Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". PNAS **102**, 13950-13955

- Tindall, B. J., Rossello-Mora, R., Busse, H.-J., Ludwig, W. and Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.* **60**, 249-266
- Top, E. M. and Springael, D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr. Opin Biotechnol*. **14**, 262-269.
- Uchiyama, I., Higuchi, T. and Kawai, M. (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Research* **38** (suppl 1), D361-D365
- Uchiumi, T., Ohwada, T., Itakura, M., Mitsui, H., Nukui, N., Dawadi, P., Kaneko, T., Tabata, S., Yokoyama, T., Tejima, K., Saeki, K., Omori, H., Hayashi, M., Maekawa, T., Sriprang, R., Murooka, Y., Tajima, S., Simomura, K., Nomura, M., Suzuki, A., Shimoda, Y., Sioya, K., Abe, M. and Minamisawa, K. (2004) Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. *J. Bacteriol.* **186**, 2439-2448.
- Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P. and Merkl. R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* **16**, 142.
- Zerbino, D.R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821-829.