Review

Mining the genomes of plant pathogenic bacteria: how not to drown in gigabases of sequence

BORIS A. VINATZER* AND SHUANGCHUN YAN

Department of Plant Pathology, Physiology, and Weed Sciences, Virginia Tech, Blacksburg, VA 24061, USA

SUMMARY

Hundreds of bacterial genomes including the genomes of dozens of plant pathogenic bacteria have been sequenced. These genomes represent an invaluable resource for molecular plant pathologists. In this review, we describe different approaches that can be used for mining bacterial genome sequences and examples of how some of these approaches have been used to analyse plant pathogen genomes so far. We review how genomes can be mined one by one and how comparative genomics of closely related genomes releases the true power of genomics. Databases and tools useful for genome mining that are publicly accessible on the Internet are also described. Finally, the need for new databases and tools to efficiently mine today's plant pathogen genomes and hundreds more in the near future is discussed.

INTRODUCTION

For bacterial plant pathogens, the genomics era officially began with the publication of the genome sequence of *Xylella fastidiosa*, the causative agent of citrus variegated chlorosis (Simpson *et al.*, 2000). Genome sequences of representative strains of most taxonomic groups that contain bacterial plant pathogens of worldwide importance soon followed. Recently, multiple species or subspecies within the same taxonomic group were added to the list of completely sequenced genomes and we have now reached a point at which several isolates of the same species are being sequenced revealing a surprisingly high degree of variability even within species. To get the latest status of bacterial plant pathogen genome sequences, the Genomes OnLine Database

(GOLD) at www.genomesonline.org/ can be consulted for a frequently updated list of genome sequencing projects.

The release of the *Xylella fastidiosa* genome represented not only the beginning of the genomics era of bacterial plant pathogens but also the beginning of the post-genomics era, which can be defined as the field of research concerned with correlating the genome sequence of an organism with its phenotype. To accomplish this task, genome sequences need to be annotated first, i.e. they need to be searched for putative genes and the function of the proteins encoded by these genes needs to be predicted. Annotated genomes are then mined for candidate genes that might be at the basis of a phenotype of interest: in the case of plant pathogens, any aspect of their interactions with plants. Besides being analysed one by one, genomes can also be compared with genomes of closely related organisms. We describe both approaches and show how they reveal synergistic effects when combined. Transcriptome and proteome analysis and molecular genetic screens can further elucidate gene function. Finally, annotation, genome comparison data and experimental data can be stored in genome databases and made accessible and searchable over the internet to maximize their utility to the research community. The review concludes with a speculation on future needs to mine hundreds of bacterial plant pathogen genome sequences, which can be expected within the next few years because of the recent dramatic reduction in DNA sequencing costs. We describe how Grid technology will probably play a central role in accomplishing this tremendous task.

We caution that we have tried to give an as complete as possible overview of genome mining approaches and tools and how they can be applied to bacterial plant pathogens, but we are aware of the fact that we were not able to cover all of them as exhaustively as they would deserve. Also, because our focus for this review was to describe genome-mining approaches, we have only referenced a limited number of results obtained by mining bacterial plant pathogen genomes. We refer the reader to Preston *et al.* (2007), which complements this review by giving an excellent overview of pathogenicity and virulence mechanisms of plant pathogenic bacteria from a genomics perspective.

*Correspondence: Tel.: +1 540 231 2126; Fax: +1 540 231 3347;
E-mail: vinatzer@vt.edu

ANNOTATION, THE FIRST STEP IN MINING A GENOME

The main goal when annotating a bacterial genome is to locate protein-coding sequences (CDSs) and assigning a putative function to these CDSs. First, all open reading frames (ORFs) that possibly code for proteins are identified. This is either done by considering all ORFs that are longer than an arbitrarily chosen length to be putative CDSs or by using interpolated Markov models (IMMs) that find ORFs likely to be CDSs based on sequence characteristics using a program like GLIMMER (Salzberg *et al.*, 1998). The proteins encoded by each putative CDSs are then compared with the proteins encoded by all genes in closely related annotated genomes using BLASTP (Altschul *et al.*, 1997). It is assumed that their functions are the same or similar to the functions of the most similar annotated gene products in the closely related genomes. All CDSs, for which no orthologue with a significant score can be found, are then compared with a non-redundant protein database like the 'nr' protein database at the National Center for Biotechnology Information (NCBI) to find similar genes in any organism. If a CDSs has as best hit a protein with unknown function, the gene product is annotated as a 'conserved hypothetical protein', whereas if there is no hit with a significant score, the gene product is annotated as a 'hypothetical protein'.

To improve the prediction of gene function, the TIGRFAM (Selengut *et al.*, 2007) and Pfam (Finn *et al.*, 2006) databases were established. These manually curated databases are collections of protein families featuring multiple sequence alignments and hidden Markov models (HMMs) describing protein motifs as statistical models. TIGRFAM and Pfam database searches are performed using the program HMMER (selab.janelia.org/) (Eddy, 1998), which provides the likelihood that a protein sequence fits the statistical model of a motif associated with a protein family compared with a random model to indicate the relative likelihood that a sequence belongs to a certain protein family (Birney, 2001). In this way, gene function is not only predicted based on the overall similarity to a protein with known function but it is based on the presence of a motif (or motifs) known to be necessary for that function. To aid in function assignment, NCBI also assembled clusters of orthologous groups (COGs) (Tatusov *et al.*, 2003), each consisting of at least three reciprocal best BLAST hits in three distantly related genomes.

After putative functions are assigned to putative CDSs, cleanup is required. In fact, the same DNA sequence may contain several putative CDSs in different reading frames and on opposite strands. By comparing BLAST scores and other evidence for function, only the CDSs most likely to represent functional genes are maintained as annotated genes. Another challenge is to choose the correct start codon of genes. The start codon can either be assumed to be the codon, which immediately precedes the region of the ORF

that aligns with its orthologous gene (or genes) in another genome, or it can be predicted based on the presence of a ribosome binding site using programs such as 'RBSfinder' (Suzek *et al.*, 2001).

The steps described above can be handled with little human intervention and are thus part of what is considered 'automated' or 'machine' annotation. Programs to perform each step have been assembled into annotation 'pipelines', e.g. the 'Annotation Engine' developed by TIGR (www.tigr.org/AnnotationEngine) or the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) developed by NCBI (www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html), to which users can submit their genome sequences for automated annotation. Automated annotation is often followed by 'community annotation', during which a community of researchers integrates the automated annotation with experimentally validated data. In the case of *Agrobacterium tumefaciens* a website was specifically designed for this purpose: agro.vbi.vt.edu/public. An extremely useful and widely used tool to integrate manual annotation with automated annotation is the program Artemis developed by the Sanger Institute (Berriman and Rutherford, 2003). Artemis allows us to display all kind of sequence analysis results and to edit them manually. For example, genes missed by a gene prediction program can be added, start codons can be changed, or additional gene names and annotator's comments can be included. The final results can then be saved for submission to GenBank.

USING A COMMON VOCABULARY TO DESCRIBE GENE FUNCTION

A very important aspect of annotation and genome mining in general is the use of a common terminology to describe gene function. Standardized terms need to be used to describe function in all genomes so that genomes can easily be queried and compared with each other. The Gene Ontology (GO) Consortium was initiated with the aim of writing a dictionary of such standardized terms that could be used in the annotation of all genomes and could facilitate comparative genomics (Consortium, 2001). Each gene function is described in relation to the associated biological processes, cellular components and molecular functions. GO terms are intended to describe functions present in all organisms independent of the species to which the organism belongs. Functions of plant pathogen genes specific to plant-pathogen interactions, i.e. the functions plant pathologists are most interested in, were not emphasized in the initial manifestation of the GO. The PAMGO (Plant Associated Microbe Gene Ontology) Consortium (pamgo.vbi.vt.edu/) filled this need by developing standardized terms for describing function of gene products involved in microbe-plant interactions. The first 472 terms were incorporated into the GO in January 2007 and can now be used to describe the function of genes in relation to their role in plant-microbe interactions.

MINING ONE GENOME AT THE TIME

The simplest, and probably most common, way of mining a genome is to do a similarity search using any flavour of BLAST (BLASTP, BLASTN, tBLASTN, or BLASTX) (Altschul *et al.*, 1997). In this case, a gene or protein sequence of known function or a gene identified in a molecular genetic screen is used as query to search for an orthologue in a genome sequence. If a protein family is conserved across different species, a protein with a known role in virulence in one plant pathogen (or animal or human pathogen) can be used to search for a similar protein in a plant pathogen genome sequence. The problem with similarity searches is that it is limited to families of proteins which members are conserved enough to be identified by the BLAST algorithm. Caution must also be used since members of the same protein family may have different functions or play different roles in different organisms. Therefore, assigned function based on comparison with other family members is always only putative until the function has been validated experimentally.

BLAST searches are not sufficient to mine genomes for the most intensively studied class of bacterial plant pathogen proteins: effector proteins injected by bacterial pathogens into host cells using the Type III secretion system (T3SS) (Galan and Wolf-Watz, 2006). Effectors are used by plant pathogens to suppress plant defences and generally to manipulate plant metabolism to create an environment favourable to pathogen growth (Grant *et al.*, 2006). The first effectors that were identified in the pre-genomic era were effectors with avirulence activity on model plant species. Avirulence effectors have virulence activity on susceptible plants but they are detected by the plant immune system on resistant plants (Alfano and Collmer, 2004). Upon detection, resistant plants activate defences to block pathogen growth. The defence response often includes programmed cell death, macroscopically visible as total leaf collapse when a high concentration of pathogen is inoculated. This plant defence response is called the hypersensitive response (HR) (Greenberg and Yao, 2004). The first effectors were identified in *Pseudomonas syringae*, a species comprised of bacterial speck, blight and canker pathogens of many crops and ornamentals, and in *Xanthomonas*, a genus that includes species that similarly to *P. syringae* cause bacterial diseases of over 100 different plant species. The avirulence gene *avrPto1* with avirulence activity on tomato is an example of one of the first isolated effectors (Ronald *et al.*, 1992). Until recently, bacterial plant pathogen effectors without avirulence activity were extremely difficult to identify since mutations in effector genes usually have weak or no virulence phenotypes. Also, while effector proteins of many bacterial animal pathogens are secreted in culture in large quantities and can be easily sequenced, bacterial plant pathogens secrete effector proteins in culture in small quantities that are difficult to analyse. Only shortly before the first genome sequences of plant pathogenic bacteria became

available, a first molecular genetic screen for effectors was successful in detecting effectors independently of their avirulence activity in *P. syringae* by using the HR eliciting activity of the C-terminus of the well-studied effector AvrRpt2 as a reporter (Guttman *et al.*, 2002). However, even this screen did not allow identification of the entire effector repertoire of the analysed strain.

With the availability of plant pathogen genome sequences mining for effector genes evolved into a real 'effector rush'. In *P. syringae*, several groups in parallel started mining available genome sequences at various stages of assembly or used genome sequences to obtain the sequence of putative effector genes identified in functional screens (Boch *et al.*, 2002; Fouts *et al.*, 2002; Guttman *et al.*, 2002; Petnicki-Ocwieja *et al.*, 2002; Zwiesler-Vollick *et al.*, 2002). Similar searches soon started in other bacterial genomes, for example in *Xanthomonas campestris* pv. *campestris*, the causative agent of black rot disease in members of the Brassicaceae family, in the bacterial citrus canker pathogen *Xanthomonas axonopodis* pv. *citri* (da Silva *et al.*, 2002), in *Xanthomonas campestris* pv. *vesicatoria* (Buttner *et al.*, 2003), which causes bacterial spot of tomato and pepper, in the bacterial wilt pathogen *Ralstonia solanacearum* (Cunnac *et al.*, 2004; Salanoubat *et al.*, 2002), and in the soft-rotting bacterium *Erwinia carotovora* ssp. *atroseptica* (Bell *et al.*, 2004). Using 'effector mining' as an example, we will now describe several different genome-mining approaches.

Mining by protein motif or pattern

Even when proteins do not belong to the same protein family, they may still share some characteristics with other proteins. Therefore, it is sometimes possible to identify a common motif or pattern in a subset of known members of a group of proteins, for example membrane proteins, and identify unknown members of the same group based on that motif or pattern. By 'motif' we mean a well-defined stretch of protein sequence with conserved amino acids at conserved positions. By 'pattern' we simply mean that amino acids with certain characteristics occur more frequently than others in a region of a protein.

In the case of effectors secreted by the T3SS of *P. syringae*, it was first noticed that the 50 most N-terminal amino acids (representing the putative secretion region that targets effectors in the bacterial cell to the T3SS) in all known effectors contained at least 10% serine, but no aspartate (Guttman and Greenberg, 2001). Searching the draft genome sequence of *P. syringae* pv. *tomato* DC3000 for any ORFs coding for proteins with that pattern, it was possible to identify a group of putative effectors. This search required nothing else than a short custom script written in the programming language Perl (see www.bioperl.org for a description of how Perl is used in bioinformatics). The putative effector genes were then further analysed by searching among

them for genes preceded by the *hrp* box promoter element, which is the typical promoter of effectors (Lan *et al.*, 2006). This led to the identification of dozens of putative effectors, most of which have been confirmed since then (Lindeberg *et al.*, 2006).

In parallel, T. Petnicki-Ocwieja and co-workers found that *P. syringae* effectors also had the following pattern: solvent exposed equivalent amino acids in the first five positions, absence of acidic residues in the first 12 positions, and an overall amphipathicity and richness in polar amino acids in the 50 most N-terminal amino acids (Petnicki-Ocwieja *et al.*, 2002). Searching the draft genome sequence of *PtoDC3000* for any ORFs coding for protein products at least 150 amino acids long with the above pattern using an HMM, they identified 400 ORFs. After manually editing the hits and restricting the hits to ORFs preceded by the conserved *hrp* box promoter element (similar to the approach used by Guttman and co-workers), T. Petnicki-Ocwieja *et al.* were left with 32 new effector candidates. Most of these candidates were identical to the effectors identified by Guttman and co-workers and have in the meantime been confirmed (Lindeberg *et al.*, 2006). These two examples show how already a draft genome sequence is sufficient to efficiently mine for a group of proteins of interest.

It is also possible to search for more restricted subgroups of effector proteins with predicted enzymatic function inside the host cell. Abramovitch and co-workers found that the *P. syringae* effector AvrPtoB (also called hopAB2_{PtoDC3000}) mimics an E3 ubiquitin ligase (Abramovitch *et al.*, 2006). It is now known that several plant and animal pathogens inject proteins containing the ubiquitin ligase motifs F-box and U-box into their hosts to interfere specifically with the proteasome (Angot *et al.*, 2006). To find yet unknown effectors with these motifs in any of the sequenced plant and animal pathogens, Angot *et al.* (2007) searched the entire microbial protein database for these motifs and identified several putative proteins in bacterial plant pathogen genomes that might interfere with the host proteasome. A more general approach was used in *Ralstonia solanacearum* (Salanoubat *et al.*, 2002). A large number of putative T3SS effectors were identified by looking for any protein with homology to plant proteins or with predicted function typical of eukaryotic proteins. These results show that mining for virulence genes by motif and pattern searches is an extremely rewarding approach.

Mining for regulatory elements

As plant pathogens have to coordinately activate their virulence mechanisms during plant infection, virulence genes can be expected to have common promoter elements activated by a common transcription factor in eukaryotes or a common sigma factor or transcriptional activator in bacteria. In *P. syringae* and in the fire blight pathogen *Erwinia amylovora*, effector genes and genes coding for T3SS structural components are preceded by the

hrp box promoter element. The alternative sigma factor HrpL is up-regulated during infection and expression of the T3SS components and effector genes is induced upon binding of HrpL to their *hrp* boxes (Lan *et al.*, 2006). In *Xanthomonas* effector genes and genes of the T3SS are preceded by the *PIP* box promoter element (Koebnik *et al.*, 2006) and in *R. solanacearum* by the *hrpII* box (Occhialini *et al.*, 2005). The search for genes downstream of *hrp* boxes was used for the first time by Fouts *et al.* (2002) to find putative effector genes. They developed an HMM of the *hrp* box, which was recently improved by Ferreira *et al.* (2006). Using the same HMM, Vencato *et al.* (2006) also found T3SS-independent virulence genes downstream of *hrp* boxes in the *P. syringae* genomes of *Pph1448A* and *PtoDC3000* and experimentally confirmed a role in virulence for these genes. Co-regulation of T3SS-dependent and -independent genes was recently also described in *R. solanacearum* (Valls *et al.*, 2006). In the genome sequence of *Xanthomonas campestris* pv. *vesicatoria* *PIP* boxes were found upstream of known and putative T3SS effectors and upstream of a gene coding for a hypothetical polygalacturonase and other T3SS-independent genes (Thieme *et al.*, 2005), suggesting that co-regulation of T3SS-dependent and -independent genes is common in bacterial plant pathogens.

Mining for anomalous regions

A bacterial genome can be divided into a core genome and a flexible genome. The core genome is vertically inherited from a common ancestor by all members of a bacterial species. Consequently, it is shared by all isolates of the species. The flexible genome consists of regions, called genomic islands (GEIs), which were acquired horizontally by members of a species. There is abundant evidence suggesting that GEIs have their origin in unrelated bacterial genomes or in bacteriophages (Abby and Daubin, 2007). GEIs of pathogens often contain genes involved in pathogenicity and are called pathogenicity islands (PAIs) (Dobrindt *et al.*, 2004). PAIs contribute to host range and virulence of plant pathogen isolates (Arnold *et al.*, 2003). Thus, mining for GEIs or PAIs is another way to identify putative virulence genes. There are two independent approaches to do this: one is based on the different sequence characteristics typical of core and flexible genomes, while the other is based on genome comparison in which GEI candidates are identified as genomic regions that are missing in one or more genomes of a species (see 'Mining by comparing two or multiple genomes' below).

Several programs have been developed to distinguish genomic islands from core genomes based on their different sequence characteristics. These programs search a genome for regions with characteristics that are typical of genomic islands: lower G+C content than the core genome, codon usage different from the core genome, flanking tRNA genes, and presence of genes indicative of horizontal transfer (Dobrindt *et al.*, 2004). Some

websites allow users to view precomputed locations of genomic islands for sequenced genomes, for example the IslandPath website (www.pathogenomics.sfu.ca/islandpath/), while other programs can be downloaded and be used to predict genomic islands in any genome sequence. A recent addition to these latter programs is Alien Hunter, which predicts genomic islands using a two-state HMM exploiting compositional biases of genomic islands (Vernikos and Parkhill, 2006).

MINING BY COMPARING TWO OR MULTIPLE CLOSELY RELATED GENOMES

So far we have described approaches focused on single genomes, yet the real power of genomics is released when genomes of a group of closely related organisms with different phenotypes are compared. In this way, phenotypic differences between organisms can be correlated with genomic differences. The closer the organisms are related, the more similar their genomes are, and the easier it will be to identify the genomic differences that determine their phenotypic differences. Comparing a pathogen with a closely related non-pathogen can, for example, help to identify the molecular basis of pathogenicity. Such a comparison is being done for the Gram-positive bacterium *Streptomyces scabies*, the causative agent of potato scab, and *Streptomyces coelicolor*, a closely related non-pathogenic soil bacterium (Loria *et al.*, 2006). Surprisingly, one unique large PAI that contains several virulence genes appears to be the most important difference between the two species and is in some cases sufficient to enable non-pathogenic *Streptomyces* species to cause potato scab in controlled conditions. Another extremely informative comparison is that between closely related plant and animal pathogens to identify the molecular basis of the ability to cause either animal or plant diseases. The family Enterobacteriaceae constitutes a group well suited for this purpose given that sequenced Enterobacteriaceae encompass strains of the human pathogens *Escherichia coli*, *Shigella*, *Salmonella* and *Yersinia* and plant pathogens of the *Erwinia* group (for example the apple blight pathogen *Erwinia amylovora* and the necrotrophic broad host range pathogen *Erwinia carotovora* ssp. *atroseptica*). Toth *et al.* (2006) identified 17 major PAIs in *Erwinia carotovora* ssp. *atroseptica* absent in the human pathogens of the family Enterobacteriaceae. The functions of the genes on these PAIs are now being experimentally dissected to determine the molecular basis of the ability of *Erwinia carotovora* ssp. *atroseptica* to cause bacterial soft rot in plants.

Comparing isolates with different host ranges of the same species, host specificity can be dissected. However, it is important to note that plant pathogenic bacteria with the same species name are not necessarily closely related. For example, although the three sequenced *P. syringae* strains have all the same species name, they are almost as different as *E. coli* and *Salmonella*

typhimurium. In fact, over one-quarter of the *P. syringae* genome is not shared between strains based on a multiple genome alignment (our unpublished data). It is hence very difficult to correlate genomic differences with phenotypic differences when comparing these strains. Instead, a very informative comparison was made between the sequenced *Xanthomonas campestris* pv. *campestris* strains 8004 (Qian *et al.*, 2005) and ATCC 33913 (da Silva *et al.*, 2002), two strains that have different host ranges but are very closely related. Only 13% of their genomes do not align. For the genes that differ between the two strains, Qian *et al.* (2005) were able to confirm experimentally that several of them play a role in host range.

We will now describe various comparative genomic approaches and list examples from analyses of animal or human pathogens when examples of such approaches have not yet been published for bacterial plant pathogens.

Mining by presence/absence of genes

As described above, to determine if a genome sequence contains a certain gene of another genome, one can take that gene and run a similarity search using the BLAST algorithm. To run a BLAST search manually with all genes of one genome as query and all genes of a second genome as subject is not practical. OrthoMCL (Li *et al.*, 2003) is one of the programs that can do this automatically. It takes all proteins of multiple genomes, blasts all proteins against each other, and finally outputs orthologous groups, each of which consists of genes that can be assumed to have evolved from a common ancestor and thus have the same or similar functions. Genes that are present only once in one of the analysed genomes are not included in any orthologous group. Therefore, determining orthologous groups in a group of genomes is very useful to rapidly identify differences in gene content between genomes and find candidate genes that may determine the phenotypic differences that exist between the organisms to which the genomes belong. Pre-computed OrthoMCL groups can be searched at orthomcl.cbil.upenn.edu for various microbial and eukaryotic genomes and the tool 'Phylogenetic Profiler' at img.jgi.doe.gov allows users easily to search orthologous groups in nearly all sequenced bacterial genomes to identify genes present or absent in any combination of multiple genomes.

Mining by presence/absence of genomic regions

We mentioned above the possibility of identifying a bacterial core genome—which is vertically inherited—from the flexible genome—which is horizontally inherited from other organisms—by their different sequence characteristics. Use of programs that perform pair-wise or multiple genome alignments makes it possible to identify all genomic regions that make up the core genome of a group of sequenced strains based on the fact that

core genome regions are present in all analysed strains and can thus be aligned with each other. Consequently, the regions of the flexible genome are those regions that are missing in at least one of the analysed strains and thus cannot be aligned with each other. These regions should not be called 'genomic islands' as they also include regions that were vertically inherited from a common ancestor of the analysed strains, but that were lost from one of the strains. The reverse can also occur: a genomic island may have been acquired by all analysed strains and may have become part of the 'core genome' based on a multiple genome alignment. The regions that are not shared between genomes should thus more generally be called 'loops' (Ohnishi *et al.*, 2002) expressing simply the fact that they cannot be aligned between genomes. Consequently, the core genome determined by comparative genomics is not the same as the core genome identified by sequence analysis of a single genome.

As different regions of the 'core' genome have different levels of conservation and several loops may contain genes of the same gene families, the main challenge one faces when running an alignment program is the choice of parameters that correctly identify most core genome regions and most flexible genome regions. Although it is impossible to use parameters that allow the correct distinction between all regions of the core and the flexible genome, one can choose the relatively best parameters using known core and flexible genome regions for calibration. The second challenge is to identify genomic rearrangements between genomes. Only two programs can currently handle genomic rearrangements: MUMmer (Kurtz *et al.*, 2004) for pair-wise alignments and MAUVE (Darling *et al.*, 2004) for pair-wise and multiple genome alignments. Whereas the program MUMmer has been applied to several genomes of bacterial plant pathogens, MAUVE analyses of bacterial plant pathogens have not yet been published.

Alignment programs should not be confused with genome visualization tools. Whereas alignment programs identify the coordinates of aligned and unaligned regions and output data as graphs and as tables, the Artemis Comparison tool (ACT) (Carver *et al.*, 2005) is a visualization tool that runs a BLAST search to identify regions with homology and represents those regions only graphically. ACT is thus a useful tool to compare specific genomes visually, but its output is not well suited for further analysis.

MINING BY COMBINING ORTHOLOGOUS SEARCHES WITH MULTIPLE GENOME ALIGNMENTS

While determining orthologous groups and genome alignments are powerful approaches, their combination releases surprisingly synergistic effects. Their combination makes it possible to identify gene loss events due to non-sense mutations, annotation errors and sequencing errors. For example, if a region can be aligned

between two genomes and this region contains one gene in one genome with no orthologous gene in the second genome, the gene in the second genome might either have been lost due to a non-sense mutation, might contain a sequencing error that created a frameshift or the gene was missed in the annotation. Therefore, by combining orthologous searches with genome alignments, sequencing errors can be identified and corrected and annotation can be improved. As effector genes with avirulence activity are frequently lost from genomes because of evolution in the host plant of a corresponding resistance gene, this kind of analysis becomes very important when comparing effector gene repertoires between genomes.

Mining by comparative evolutionary genomics

Homologous recombination, i.e. the replacement of a genomic region in one bacterium with a homologous region acquired from another member of the same species, has been found to be an important evolutionary mechanism in some species, while other species have been found to be largely clonal, i.e. they very rarely recombine (Narra and Ochman, 2006). Often, recombination is assessed by multi-locus sequence typing (MLST) (Maiden, 2006), i.e. sequencing 5–7 gene fragments in several strains of a species and analysing the aligned sequences with a series of statistical tests either to identify recombination breakpoints or to measure the contribution of recombination to variation between strains. Identifying recombination at the whole genome level has become feasible with the availability of genome sequences of several strains of the same species. So far, recombination has not been determined at the whole genome level in any plant pathogenic bacterium, but Mau *et al.* (2006) have used the multiple genome alignment program MAUVE to identify recombination breakpoints and genomic regions with aberrant phylogeny using six *E. coli* and *Shigella flexneri* genomes. They found that recombination is an important factor in the evolution of *E. coli*.

Another important evolutionary parameter is the selection pressure genes are under. Genes under purifying selection have a non-synonymous/synonymous mutation rate ratio $\omega = dN/dS$ close to zero to maintain function by conserving their amino acid sequence. Examples are housekeeping genes that function in basic metabolic pathways. Genes under neutral selection have a dN/dS ratio close to 1. Examples are housekeeping genes that can change their protein sequence without losing their function. Genes under positive selection have a dN/dS ratio larger than 1 as they are under selection pressure to change their sequence; for example, antigens of animal pathogens or some effectors of plant pathogens are under selection pressure to change their protein sequence to avoid recognition by the host immune system. Determining the selection pressure genes are under can hence give important hints to which genes in a plant pathogen genome evolve in response to the host immune system

and are thus important for virulence. This can be done by applying programs that determine selection pressure, such as CODEML of the PAML package (Yang, 1997), to all orthologous groups identified in a group of closely related genomes.

Phylogenomics is a particular comparative evolutionary approach used to predict the function of proteins in a genome. Instead of simply using BLAST to identify orthologous groups and then predicting the function of a gene in a new genome based on the function of a protein in the same orthologous group in an annotated genome, phylogenomics includes phylogenetic analysis of that orthologous group (Eisen, 1998; Sjolander, 2004). For example, if a gene in an orthologous group comprises genes coding for different kinases, it is not easy to determine which kinase the gene encodes. Determining the position of the gene in a phylogenetic tree may instead clearly show that the gene is located on a branch derived from an ancestor with a specific kinase activity and based on this location in the tree the most likely kinase activity can be inferred.

MINING GENOMES THROUGH EXPERIMENTS

Genome sequences can be used in two kinds of experiments: experiments that are facilitated by a genome sequence and experiments that require a genome sequence.

There are many examples of the first kind of experiment. Molecular genetic screens are just as easy to perform in an unsequenced organism as in a sequenced organism. For example, genes in a genome can be disrupted by transposon-mutagenesis and resulting mutants can be screened for a phenotype. The difference is that in an unsequenced genome it is very labour-intensive to sequence the regions flanking the transposon insertion sites to obtain the entire sequence of the mutated genes. If a sequenced genome is used instead, all that is needed are 50 bp of sequence flanking the insertion to identify the gene in the genome sequence that was mutated. This can be done easily and cheaply by Arbitrary Primed (AP) PCR (Jacobs *et al.*, 2003), which makes it possible for one person to determine sequences flanking hundreds of transposon insertions in one day. For example, AP PCR was used in a screen to identify regulators of the T3SS in the sequenced biocontrol strain *P. fluorescens* SBW25 (Jackson *et al.*, 2005). *P. fluorescens* SBW25 was first mutated with a mini transposon, the resulting mutant library was then screened for increased expression of the T3SS, and the transposon insertions were finally identified in those transposon insertion lines that showed increased expression of the T3SS using AP PCR. In *E. carotovora* ssp. *atroseptica*, by contrast, the genome sequence was used as a starting point to identify genes with a role in virulence: first, a mini transposon library was constructed and insertions in candidate virulence genes were selected; second, transposon-insertion lines were used in plant infections to find among the selected genes those that in fact have a virulence function (Bell *et al.*, 2004).

To restrict the number of genes of a plant pathogen genome that may play a role in virulence, one can determine the genes that are transcribed during plant infection compared with the genes transcribed during growth of the same pathogen in a rich culture medium. Microarrays spotted with oligonucleotides corresponding to all predicted CDS or even the entire pathogen genome can be used for this purpose. Justin Borevitz and colleagues developed an Affymetrix oligoarray for *Arabidopsis thaliana* that also contains the sequenced *P. syringae* genomes and sequences from other plant pathogens, which makes it possible to determine expression of *P. syringae* genes and *A. thaliana* genes during the same infection (Zhang *et al.*, 2007). However, it is difficult to extract RNA from a pathogen during plant infection as pathogen RNA presents only a relatively small portion of the total RNA extracted from infected leaves. While there have been reports of successful gene expression studies of pathogens during infection, for example of *Blumeria graminis* f. sp. *Hordei* during infection of barley (Both *et al.*, 2005) or of *Phytophthora sojae* during infection of soybean (Torto-Alalibo *et al.*, 2007), in the case of bacterial pathogens there is the additional complication of dealing with mRNA without polyA tail. This makes it difficult to separate mRNA from ribosomal RNA. Fortunately, some pathogens can be grown in medium that resembles the conditions during plant infection and RNA can be extracted from pathogens grown in that medium. This has been done, for example, by growing *P. syringae* in minimum medium, which is known to activate the HrpL regulon (Ferreira *et al.*, 2006). In the case of *P. syringae*, a 'sleight of hand' has been used: as HrpL is known to regulate expression of effector genes, a *P. syringae* strain that constitutively expresses HrpL was engineered and gene expression was compared between wild-type *P. syringae* and the HrpL overexpresser line, both grown in rich medium (Fouts *et al.*, 2002). This allowed the identification of a series of genes regulated by HrpL that are putatively expressed during infection and have a putative role in virulence.

Determining gene expression during infection is not only a way of mining for genes that may play an important role in virulence. Whole genome expression analysis is also very important to determine which of the hypothetical genes annotated in a genome are actually expressed. These data can then flow back into the annotation of the genome so that a 'hypothetical protein' can be annotated as a 'gene transcribed during infection'.

While microarrays are an invaluable tool to determine genome-wide gene expression, it is impossible to use microarrays to detect very low expression of genes. A more sensitive approach is to sequence a sample of the mRNA population of a pathogen during infection, for example using 454 pyrosequencing technology (Margulies *et al.*, 2005). The GS FLX sequencer (Roche, Switzerland), based on 454 technology, has a sequencing capacity of 400 000–500 000 DNA molecules per run with an average read length of 250 bp. Therefore, pyrosequencing has the potential to detect

most bacterial genes transcribed during infection as long as rRNA is efficiently removed from the extracted total RNA before reverse transcription. If short cDNA fragments are ligated together into longer fragments and then sequenced, as in 'Serial Analysis of Gene Expression' (SAGE) (Velculescu *et al.*, 1995), fragments of several million transcripts can be assayed in one 454 sequencing run. SAGE has already been used to study gene expression in pathogens and hosts simultaneously using Sanger sequencing (Jantasuriyarat *et al.*, 2005; Matsumura *et al.*, 2003) and we can expect to see many more such studies using pyrosequencing technology. Also, SOLEXA sequencing (Bentley, 2006) is ideally suited for transcriptome analysis. Millions of DNA molecules, each 25 bp long, can be sequenced with this technology at an even lower cost than that of 454 technology.

Another powerful approach of mining for genes that may be important for virulence is to determine which genes in a genome are not only transcribed, but also expressed, by combining genomics with proteomics. Even more than with transcriptome analysis, the major difficulty in proteomics is to obtain enough pathogen proteins from infected plants. Because of this problem, proteomics has mainly been applied to the analysis of plant responses to pathogens. For example, Jones *et al.* (2006) identified proteins up-regulated in *A. thaliana* during *P. syringae* infection. Again, once the proteome of a pathogen has been determined, this information can flow back into the genome annotation and a 'hypothetical protein' can become an 'expressed protein' or a 'protein expressed during infection'.

INTEGRATED MINING

After a group of closely related genomes has been sequenced and annotated, anomalous regions have been identified, genomes have been compared with each other, and genomes and genes have been subject to an evolutionary analysis, all these results can be combined in one relational database. If the transcriptome and/or proteome of a pathogen has been determined in different conditions and the mutant phenotype of genes has been analysed on a genome-wide scale, this information can also be added. The database ASAP (Glasner *et al.*, 2006) is an excellent example of an integrated database that contains annotation and experimental data and allows the comparison of genomes by viewing orthologous groups. However, we envisage future integrated databases to go one step further by making it possible to search genomes for genes using any combination of multiple parameters. This would make it possible, for example, to search for predicted kinases, present in a subset of strains of a species, under positive selection, and expressed during infection.

WEB RESOURCES FOR DATA MINING

Web resources for data mining can be divided into two groups: database websites and tool websites. A selection of these websites with their URL and with short descriptions are listed in Table 1 and Table 2, respectively. More databases can be found

Table 1 Database websites for genome mining.

Name & URL	Description	Reference
ASAP asap.ahabs.wisc.edu/asap/home.php	<ul style="list-style-type: none"> A relational database and web interface developed to store, update and distribute genome sequence data and gene expression data collected by or in collaboration with researchers at the University of Wisconsin-Madison 	(Glasner <i>et al.</i> , 2006)
Bacterial genome sequences at the Sanger Institute www.sanger.ac.uk/Projects/Microbes/	<ul style="list-style-type: none"> View, search, compare and download genomes sequenced at the Sanger Institute including plant pathogen genomes 	
CBS Genome Atlas www.cbs.dtu.dk/services/GenomeAtlas	<ul style="list-style-type: none"> Downloadable genome atlases in pdf format: Genome Atlas, Structure Atlas, Base Atlas, Cruciform Atlas, A-DNA Atlas, Z-DNA Atlas 	(Hallin and Ussery, 2004)
Genome Sequencing of Agrobacterium Biovar Type Strains agro.vbi.vt.edu/public/	<ul style="list-style-type: none"> Search, annotate, and download Agrobacterium biovar sequences 	
JGI IMG img.jgi.doe.gov	<ul style="list-style-type: none"> Many mining tools for microbial genes and genomes similar to TIGR CMR and results can be placed in cart for download Includes most published, and also many yet unfinished and unpublished, genomes Phylogenetic Profiler allows finding genes that are shared or different between any finished or unfinished genome 	(Markowitz <i>et al.</i> , 2006)
MicrobesOnline www.microbesonline.org	<ul style="list-style-type: none"> Comparative, evolutionary, and functional microbial genomics of most published microbial genomes 	(Alm <i>et al.</i> , 2005)

Table 1 continued.

Name & URL	Description	Reference
NCBI www.ncbi.nlm.nih.gov	<ul style="list-style-type: none"> • Precomputed phylogenetic trees for all gene families and comparison with species trees • Includes protein family analyses, metabolic maps (KEGG), links to research papers, operon prediction for every genome • Users can save genes of interest, compute motifs, alignments, and phylogenetic trees, add them to a 'cart' and download them 	
	<ul style="list-style-type: none"> • Search most finished and unfinished genome sequences by homology • Entrez: text-based search and retrieval system; can be used to gather information about a single gene with one search, including gene annotation, associated literature, conserved domains within a protein, 3D structure and more • Find information on gene content and genome characteristics of most finished genome sequences • The 'BLink' link on the protein database pages gives access to precomputed taxonomic distribution for each protein 	(Wheeler <i>et al.</i> , 2007)
NCBI FTP site www.ncbi.nlm.nih.gov/Ftp/	<ul style="list-style-type: none"> • Download genome sequence and genome annotation files, programs and tools to your computer 	(Wheeler <i>et al.</i> , 2007)
NMPDR National Microbial Pathogen www.nmpdr.org	<ul style="list-style-type: none"> • Describes central and secondary metabolism, complex structures, virulence and other phenotypes in most published microbial genomes 	(McNeil <i>et al.</i> , 2007)
Data Resource	<ul style="list-style-type: none"> • Provides corrected functional annotations in a structured biological context • Flexible search terms • Comparative genomics • Interactive training materials available • is restricted to human pathogens, all other genomes are for comparative analysis 	
<i>Pseudomonas syringae</i> Genome Resources www.pseudomonas-syringae.org	<ul style="list-style-type: none"> • View and compare <i>P. syringae</i> genome sequences • View T3SS effector database 	
TIGR Comprehensive Microbial Resource cmr.tigr.org/tigr-scripts/CMR	<ul style="list-style-type: none"> • This is the most comprehensive microbial resource we could find • Search for genes by name, accession number, or role category and add search results to 'cart' for download • Search for genomes, add results to 'cart' and download • View entire genomes or genome regions • List intergenic regions • Search genomes by homology • Search proteins in any genome for a specified motif • Pseudo 2D gel computed in silico • G+C content display with customizable window size • Many comparative tools for protein and genome homology, protein scatter plot, G+C comparison graph, region comparison, genome alignments based on DNA or protein sequences with MUMmer 	(Peterson <i>et al.</i> , 2001)
	(Kurtz <i>et al.</i> , 2004)	

Table 2 Websites with tools for genome mining.

Name & URL	Description	Reference
ACT www.sanger.ac.uk/Software/ACT/	<ul style="list-style-type: none"> • Software for DNA sequence comparison and viewing based on Artemis • Visualizes genome comparisons based on BLAST 	(Carver <i>et al.</i> , 2005)
Alien_hunter www.sanger.ac.uk/Software/analysis/alien_hunter/	<ul style="list-style-type: none"> • Software to predict genomic islands in genome sequences 	(Vernikos and Parkhill, 2006)
Artemis www.sanger.ac.uk/Software/Artemis/	<ul style="list-style-type: none"> • Genome viewer and annotation tool that allows visualization of sequence features and of results from analyses within the context of the sequence 	(Rutherford <i>et al.</i> , 2000)
HMMER hmm.janelia.org/	<ul style="list-style-type: none"> • Software used to search protein databases like TIGRfam and Pfam using profile hidden Markov models 	
IslandPath www.pathogenomics.sfu.ca/islandpath	<ul style="list-style-type: none"> • Precomputed location of predicted genomic islands 	(Hsiao <i>et al.</i> , 2003)
Manatee manatee.sourceforge.net/	<ul style="list-style-type: none"> • Software tools for web-based gene evaluation and manual genome annotation • View, modify, and store annotation for prokaryotic and eukaryotic genomes 	
MAUVE gel.ahabs.wisc.edu/mauve/	<ul style="list-style-type: none"> • Software for multiple genome alignment with detailed output that allows further data processing. Gives also good graphical output of sequence alignment including genome rearrangements 	(Darling <i>et al.</i> , 2004)
MEME meme.nbcr.net	<ul style="list-style-type: none"> • Analyses sequences for similarities among them and produces a description (motif) for each pattern it discovers based on HMMs 	(Bailey <i>et al.</i> , 2006)
MUMmer mummer.sourceforge.net/	<ul style="list-style-type: none"> • Software to perform pair-wise genome alignments based on DNA sequences or protein sequences 	(Kurtz <i>et al.</i> , 2004)
NBC nbc11.biologie.uni-kl.de/	<ul style="list-style-type: none"> • Tools for genome annotation including Glimmer and RBSfinder 	
NCBI Microbial Genomes Resources www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html	<ul style="list-style-type: none"> • Provides genome annotation tools including the automated genome annotation pipeline PGAAP 	
OrthoMCL orthomcl.cbil.upenn.edu/ORTHOMCL/	<ul style="list-style-type: none"> • Software to identify orthologous groups in a set of proteins, for example all predicted proteins in two or more genomes 	(Li <i>et al.</i> , 2003)
Pfam www.sanger.ac.uk/Software/Pfam/	<ul style="list-style-type: none"> • Database of multiple sequence alignments and HMMs covering many protein domains and families • Can be searched by DNA sequence, protein sequence, protein name and more 	(Finn <i>et al.</i> , 2006)
TIGRfam www.tigr.org/TIGRFAMs/	<ul style="list-style-type: none"> • Website to search the TIGRfam database of protein families using a submitted protein sequence or name as query • Similar to Pfam database 	
WebACT www.webact.org/WebACT/home	<ul style="list-style-type: none"> • Visualizes precomputed ACT sequence comparisons 	(Abbott <i>et al.</i> , 2005)

in the 2007 Database special of *Nucleic Acids Research* and in a database and software review for the analysis and comparison of prokaryotic genomes by Field *et al.* (2005).

The most comprehensive database website that includes nearly all publicly available sequences of all organisms is housed at NCBI (see Table 1 for the NCBI URL and the URLs of the databases mentioned below). Although the NCBI database is comprehensive, three other websites have been established in the last few years that provide additional ways to mine microbial genomes sequences. These databases are: 'IMG' (Integrated Microbial Genomes) at the Joint Genome Initiative, 'CMR' (Comprehensive Microbial Resource) at TIGR, and 'microbesonline' at the Lawrence Berkeley Laboratory. All three databases allow users to select from several search parameters to search for genes, and download gene or protein sequences including a wide array of gene and protein attributes.

FUTURE NEEDS FOR MINING HUNDREDS AND THOUSANDS OF PLANT PATHOGEN DRAFT GENOMES

At the end of 2006, Roche (Basel, Switzerland) released the GS FLX sequencer, a new sequencer based on 454 pyrosequencing technology. As described above, this sequencer has average read lengths of 250 bp and can generate over 100 Mb of sequence in one run for a fraction of cost required for Sanger sequencing. The capacity of the GS FLX sequencer combined with its low cost will undoubtedly lead to the generation of draft genome sequences of hundreds of plant pathogenic bacteria within the next few years. In addition SOLEXA sequencing technology, which produces 25-bp reads totalling 1 billion sequences in one run (Bentley, 2006) can be predicted to become soon widely used to obtain sequences of organisms closely related to an already sequenced organism to find single nucleotide differences (SNPs). SNPs can then be used for pathogen identification and for comparative evolutionary studies (e.g. genome-wide recombination and selection analysis).

What does this mean for data mining? Although the cost of obtaining a draft genome sequence has dropped, the costs of complete assembly, of re-sequencing genomic regions with possible sequencing errors and of manual annotation have remained high. Therefore, mining approaches need to be tailored to machine annotated draft genome sequences that inevitably contain more sequencing and annotation errors than completely assembled manually annotated sequences. Providing quality scores for each annotated gene could be one way to address this problem. This will make it possible, for example, to determine if an interesting mutation in a known virulence gene in a draft genome sequence is more likely to be the result of a sequencing or assembly error or a true mutation. The second challenge lies in the computing power necessary to annotate, mine and compare more and more genomes. The computing resources at many individual labs or

even institutional core labs are not sufficient to perform these analyses. A solution to this problem consists in Grid-based systems that harvest the computing power of multiple computers. Such distributed computing environments (DCEs) are for example used by TIGR (www.tigr.org/grid/) to precompute the data in their microbial genome database, the TIGR-CMR (Peterson *et al.*, 2001). The Argonne National Laboratory provides users access to GNARE (GeName Analysis and Research Environment), an automated genome analysis pipeline for user-submitted genomes (Sulakhe *et al.*, 2007), and PUMA2 (Maltsev *et al.*, 2006), an interactive system for the analysis of publicly available genomes with a focus on metabolic pathways, using several Grid resources. Also, the Microbase project (Sun *et al.*, 2005) aims at developing a system that will allow users to perform integrated mining by gaining access to a computer grid through a web interface. At the heart of the planned system would lie a genome pool that is automatically updated when new genome sequences become available and a genome comparison pool that contains comparative data obtained through a suite of comparison tools. Thus far, a prototype of Microbase, called Microbase-Lite, has been developed (Sun *et al.*, 2005). In parallel, the Taverna project (Hull *et al.*, 2006) was initiated to develop a system that allows users to stack different programs written in different languages and run on different servers into a workflow and execute them using one single web interface. Behind the scenes, web service technology enables programs on the different servers to talk to each other. Integration of Taverna (or similar tools) with Microbase and the many other available databases and tools will be critical in making it possible to take full advantage of the ever growing number of genome sequences ... and not to drown in them.

ACKNOWLEDGEMENTS

Research in our laboratory is funded by Virginia Tech Start Up Funds and a Virginia Tech ASPIRES grant to B.A.V. We thank Christopher B. Lawrence (Virginia Bioinformatics Institute), Marcus C. Chibucos (Virginia Bioinformatics Institute) and Toni J. Mohr (Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech) for critical reading of the manuscript.

REFERENCES

- Abbott, J.C., Aanensen, D.M., Rutherford, K., Butcher, S. and Spratt, B.G. (2005) WebACT—an online companion for the Artemis Comparison Tool. *Bioinformatics*, **21**, 3665–3666.
- Abby, S. and Daubin, V. (2007) Comparative genomics and the evolution of prokaryotes. *Trends Microbiol.* **15**, 135–141.
- Abramovitch, R.B., Janjusevic, R., Stebbins, C.E. and Martin, G.B. (2006) Type III effector AvrPtoB requires intrinsic E3 ubiquitin ligase activity to suppress plant cell death and immunity. *Proc. Natl Acad. Sci. USA*, **103**, 2851–2856.

- Alfano, J.R. and Collmer, A. (2004) Type III secretion system effector proteins: double agents in bacterial disease and plant defense. *Annu. Rev. Phytopathol.* **42**, 385–414.
- Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L. and Arkin, A.P. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res.* **15**, 1015–1022.
- 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. *Nucleic Acids Res.* **25**, 3389–3402.
- Angot, A., Peeters, N., Lechner, E., Vailleau, F., Baud, C., Gentzbittel, L., Sartorel, E., Genschik, P., Boucher, C. and Genin, S. (2006) *Ralstonia solanacearum* requires F-box-like domain-containing type III effectors to promote disease on several host plants. *Proc. Natl Acad. Sci. USA*, **103**, 14620–14625.
- Angot, A., Vergunst, A., Genin, S. and Peeters, N. (2007) Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems. *Plos Pathog.* **3**: e3.
- Arnold, D.L., Pitman, A. and Jackson, R.W. (2003) Pathogenicity and other genomic islands in plant pathogenic bacteria. *Mol. Plant Pathol.* **4**, 407–420.
- Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34** (Web Server issue), W369–373.
- Bell, K.S., Sebaihia, M., Pritchard, L., et al. (2004) Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. *Proc. Natl Acad. Sci. USA*, **101**, 11105–11110.
- Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552.
- Berriman, M. and Rutherford, K. (2003) Viewing and annotating sequence data with Artemis. *Brief Bioinform.* **4**, 124–132.
- Birney, E. (2001) Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.* **45**, 449–454.
- Boch, J., Joardar, V., Gao, L., Robertson, T.L., Lim, M. and Kunkel, B.N. (2002) Identification of *Pseudomonas syringae* pv. *tomato* genes induced during infection of *Arabidopsis thaliana*. *Mol. Microbiol.* **44**, 73–88.
- Both, M., Csukai, M., Stumpf, M.P. and Spanu, P.D. (2005) Gene expression profiles of *Blumeria graminis* indicate dynamic changes to primary metabolism during development of an obligate biotrophic pathogen. *Plant Cell*, **17**, 2107–2122.
- Buttner, D., Noel, L., Thieme, F. and Bonas, U. (2003) Genomic approaches in *Xanthomonas campestris* pv. *vesicatoria* allow fishing for virulence genes. *J. Biotechnol.* **106**, 203–214.
- Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, **21**, 3422–3423.
- Consortium, Gene Ontology (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433.
- Cunnac, S., Boucher, C. and Genin, S. (2004) Characterization of the cis-acting regulatory element controlling HrpB-mediated activation of the type III secretion system and effector genes in *Ralstonia solanacearum*. *J. Bacteriol.* **186**, 2309–2318.
- Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403.
- Dobrindt, U., Hochhut, B., Hentschel, U. and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* **2**, 414–424.
- Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**, 163–167.
- Ferreira, A.O., Myers, C.R., Gordon, J.S., Martin, G.B., Vencato, M., Collmer, A., Wehling, M.D., Alfano, J.R., Moreno-Hagelsieb, G., Lamboy, W.F., DeClerck, G., Schneider, D.J. and Cartinhour, S.W. (2006) Whole-genome expression profiling defines the HrpL regulon of *Pseudomonas syringae* pv. *tomato* DC3000, allows de novo reconstruction of the Hrp cis element, and identifies novel coregulated genes. *Mol. Plant-Microbe Interact.* **19**, 1167–1179.
- Field, D., Feil, E.J. and Wilson, G.A. (2005) Databases and software for the comparison of prokaryotic genomes. *Microbiology*, **151**, 2125–2132.
- Finn, R.D.J., Mistry, B., Schuster-Bockler, S., Griffiths-Jones, V., Hollich, T., Lassmann, S., Moxon, M., Marshall, A., Khanna, R. and Durbin, S.R., et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.* **34** (Database issue), D247–251.
- Fouts, D.E., Abramovitch, R.B., Alfano, J.R., et al. (2002) Genomewide identification of *Pseudomonas syringae* pv. *tomato* DC3000 promoters controlled by the HrpL alternative sigma factor. *Proc. Natl Acad. Sci. USA*, **99**, 2275–2280.
- Galan, J.E. and Wolf-Watz, H. (2006) Protein delivery into eukaryotic cells by type III secretion machines. *Nature*, **444**, 567–573.
- Glasner, J.D., Rusch, M., Liss, P., Plunkett, 3rd, G., Cabot, E.L., Darling, A., Anderson, B.D., Infield-Harm, P., Gilson, C.M. and Perna, N.T. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic Acids Res.* **34** (Database issue), D41–45.
- Grant, S.R., Fisher, E.J., Chang, J.H., Mole, B.M. and Dangl, J.L. (2006) Subterfuge and manipulation: type III effector proteins of phytopathogenic bacteria. *Annu. Rev. Microbiol.* **60**, 425–449.
- Greenberg, J.T. and Yao, N. (2004) The role and regulation of programmed cell death in plant-pathogen interactions. *Cell Microbiol.* **6**, 201–211.
- Guttman, D.S. and Greenberg, J.T. (2001) Functional analysis of the type III effectors AvrRpt2 and AvrRpm1 of *Pseudomonas syringae* with the use of a single-copy genomic integration system. *Mol. Plant-Microbe Interact.* **14**, 145–155.
- Guttman, D.S., Vinatzer, B.A., Sarkar, S.F., Ranall, M.V., Kettler, G. and Greenberg, J.T. (2002) A functional screen for the type III (Hrp) secretome of the plant pathogen *Pseudomonas syringae*. *Science*, **295**, 1722–1726.
- Hallin, P.F. and Ussery, D.W. (2004) CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics*, **20**, 3682–3686.
- Hsiao, W., Wan, I., Jones, S.J. and Brinkman, F.S. (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418–420.
- Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* **34** (Web Server issue), W729–732.
- Jackson, R.W., Preston, G.M. and Rainey, P.B. (2005) Genetic characterization of *Pseudomonas fluorescens* SBW25 *rsp* gene expression in the phytosphere and in vitro. *J. Bacteriol.* **187**, 8477–8488.
- Jacobs, M.A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., Chun-Rong, L.,

- Guenther, D., Bovee, D., Olson, M.V. and Manoil, C. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA*, **100**, 14339–14344.
- Jantasuriyarat, C., Gowda, M., Haller, K., Hatfield, J., Lu, G., Stahlberg, E., Zhou, B., Li, H., Kim, H., Dean, Y.Y., R.A., Wing, R.A., Soderlund, C. and Wang, G.L. (2005) Large-scale identification of expressed sequence tags involved in rice and rice blast fungus interaction. *Plant Physiol.* **138**, 105–115.
- Jones, A.M., Thomas, V., Bennett, M.H., Mansfield, J. and Grant, M. (2006) Modifications to the Arabidopsis defense proteome occur prior to significant transcriptional change in response to inoculation with *Pseudomonas syringae*. *Plant Physiol.* **142**, 1603–1620.
- Koebnik, R., Kruger, A., Thieme, F., Urban, A. and Bonas, U. (2006) Specific binding of the *Xanthomonas campestris* pv. *vesicatoria* AraC-type transcriptional activator HrpX to plant-inducible promoter boxes. *J. Bacteriol.* **188**, 7652–7660.
- 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.
- Lan, L., Deng, X., Zhou, J. and Tang, X. (2006) Genome-wide gene expression analysis of *Pseudomonas syringae* pv. *tomato* DC3000 reveals overlapping and distinct pathways regulated by hrpL and hrpRS. *Mol. Plant-Microbe Interact.* **19**, 976–987.
- Li, L. and Stoeckert, C.J., Jr and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.
- Lindeberg, M., Cartinhour, S., Myers, C.R., Schechter, L.M., Schneider, D.J. and Collmer, A. (2006) Closing the circle on the discovery of genes encoding Hrp regulon members and type III secretion system effectors in the genomes of three model *Pseudomonas syringae* strains. *Mol. Plant-Microbe Interact.* **19**, 1151–1158.
- Loria, R., Kers, J. and Joshi, M. (2006) Evolution of plant pathogenicity in *Streptomyces*. *Annu. Rev. Phytopathol.* **44**, 469–487.
- Maiden, M.C. (2006) Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–588.
- Maltsev, N., Glass, E., Sulakhe, D., Rodriguez, A., Syed, M.H., Bompada, T., Zhang, Y. and D'Souza, M. (2006) PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.* **34** (Database issue), D369–372.
- Margulies, M.M., Egholm, W.E., Altman, S., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N. and Kyrpides, N.C. (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34** (Database issue), D344–348.
- Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D.H. and Terauchi, R. (2003) Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proc. Natl Acad. Sci. USA*, **100**, 15718–15723.
- Mau, B., Glasner, J.D., Darling, A.E. and Perna, N.T. (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol.* **7**, R44.
- McNeil, L.K., Reich, C., Aziz, R.K., et al. (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res.* **35** (Database issue), D347–353.
- Narra, H.P. and Ochman, H. (2006) of what use is sex to bacteria? *Curr. Biol.* **16**, R705–R710.
- Occhialini, A., Cunnac, S., Reymond, N., Genin, S. and Boucher, C. (2005) Genome-wide analysis of gene expression in *Ralstonia solanacearum* reveals that the hrpB gene acts as a regulatory switch controlling multiple virulence pathways. *Mol. Plant-Microbe Interact.* **18**, 938–949.
- Ohnishi, M., Terajima, J., Kurokawa, K., Nakayama, K., Murata, T., Tamura, K., Ogura, Y., Watanabe, H. and Hayashi, T. (2002) Genomic diversity of enterohemorrhagic *Escherichia coli* O157 revealed by whole genome PCR scanning. *Proc. Natl Acad. Sci. USA*, **99**, 17043–17048.
- Pallen, M.J., Nelson, K.E. and Preston, G.M. (2007) *Bacterial Pathogenomics*. Washington, DC: ASM Press.
- Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. and White, O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res.* **29**, 123–125.
- Petnicki-Ocwieja, T., Schneider, D.J., Tam, V.C., Chancey, S.T., Shan, L., Jamir, Y., Schechter, L.M., Janes, M.D., Buell, C.R., Tang, X., Collmer, A. and Alfano, J.R. (2002) Genomewide identification of proteins secreted by the Hrp type III protein secretion system of *Pseudomonas syringae* pv. *tomato* DC3000. *Proc. Natl Acad. Sci. USA*, **99**, 7652–7657.
- Preston, G.M., Guttman, D.S., Toth, I. (2007) Genomic analysis of plant pathogenic bacteria. In: *Bacterial Pathogenomics* (Pallen, M.J., Nelson, K.E., Preston, G.M., eds). pp. 392–418, Washington, DC: American Society for Microbiology.
- Qian, W., Jia, Y., Ren, S.X., et al. (2005) Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. *campestris*. *Genome Res.* **15**, 757–767.
- Ronald, P.C., Salmeron, J.M., Carland, F.M. and Staskawicz, B.J. (1992) The cloned avirulence gene avrPto induces disease resistance in tomato cultivars containing the Pto resistance gene. *J. Bacteriol.* **174**, 1604–1611.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
- Salanoubat, M., Genin, S., Artiguenave, F., et al. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*, **415**, 497–502.
- Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548.
- Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R. and White, O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.* **35** (Database issue), D260–264.
- da Silva, A.C., Ferro, J.A., Reinach, F.C., et al. (2002) Comparison of the genomes of two *Xanthomonas pathogens* with differing host specificities. *Nature*, **417**, 459–463.
- Simpson, A.J., Reinach, F.C., Arruda, P., et al. (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature*, **406**, 151–157.
- Sjolander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.
- Sulakhe, D., D'Souza, M., Syed, M., Rodriguez, A., Zhang, Y., Glass, E.M., Romine, M.F. and Maltsev, N. (2007) GNARE—a grid-based server for the analysis of user, submitted genomes. *Nucleic Acids Res.* in press.
- Sun, Y., Wipat, A., Pocock, M., Lee, P.A., Watson, P., Flanagan, K. and

- Worthington, J.T. (2005) A Grid-based System for Microbial Genome Comparison and Analysis. *Paper Read at 5th IEEE International Symposium on Cluster Computing and the Grid (Ccgird 2005), May 9–12, at Cardiff, UK.*
- Suzek, B.E., Ermolaeva, M.D., Schreiber, M. and Salzberg, S.L. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
- Thieme, F., Koebnik, R., Bekel, T., *et al.* (2005) Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris* pv. *vesicatoria* revealed by the complete genome sequence. *J. Bacteriol.* **187**, 7254–7266.
- Torto-Alalibo, T.A., Tripathy, S., Smith, B.M., Arredondo, F.D., Zhou, L., Li, H., Chibucos, M.C., Qutob, D., Gijzen, M., Mao, C., Sobral, B.W.S., Waugh, M.E., Mitchell, T.K., Dean, R.A. and Tyler, B.M. (2007) Expressed sequence tags from *Phytophthora sojae* reveal genes specific to development and infection. *Mol. Plant–Microbe Interact.* in press.
- Toth, I.K., Pritchard, L. and Birch, P.R. (2006) Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu. Rev. Phytopathol.* **44**, 305–336.
- Valls, M., Genin, S. and Boucher, C. (2006) Integrated regulation of the type III secretion system and other virulence determinants in *Ralstonia solanacearum*. *Plos Pathog.* **2**: e82.
- Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
- Vencato, M., Tian, F., Alfano, J.R., Buell, C.R., Cartinhour, S., DeClerck, G.A., Guttman, D.S., Stavrinides, J., Joardar, V., Lindeberg, M., Bronstein, P.A., Mansfield, J.W., Myers, C.R., Collmer, A. and Schneider, D.J. (2006) Bioinformatics-enabled identification of the HrpL regulon and type III secretion system effector proteins of *Pseudomonas syringae* pv. *phaseolicola* 1448A. *Mol. Plant–Microbe Interact.* **19**(11), 1193–1206.
- Vernikos, G.S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics*, **22**, 2196–2203.
- Wheeler, D.L.T., Barrett, D.A., Benson, S.H., *et al.* (2007) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35** (Database issue): D5–12.
- Yang, Z. (1997) A Program Package for Phylogenetic Analysis by Maximum Likelihood. *Comput. Applic. Biosci.* **13**, 555–556.
- Zhang, X., Richards, E.J. and Borevitz, J.O. (2007) Genetic and epigenetic dissection of cis regulatory variation. *Curr. Opin. Plant Biol.* **10**, 142–148.
- Zwiesler-Vollick, J., Plovianich-Jones, A.E., Nomura, K., Bandyopadhyay, S., Joardar, V., Kunkel, B.N. and He, S.Y. (2002) Identification of novel hrp-regulated genes through functional genomic analysis of the *Pseudomonas syringae* pv. *tomato* DC3000 genome. *Mol. Microbiol.* **45**, 1207–1218.