Chapter 22

The *Azotobacter vinelandii* Genome: An Update

João C. Setubal
Departamento de Bioquímica, Universidade de São Paulo, São Paulo, Brazil; Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Nalvo F. Almeida
Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, MS, Brazil

22.1 INTRODUCTION

This chapter presents an updated and expanded description of the *Azotobacter vinelandii* strain DJ genome. The first description of this genome was published in 2009 (Setubal et al., 2009). In that publication, the focus was on the following sets of genes: respiration and respiratory protection genes, nitrogen-fixation-related genes, genes related to other oxygen-sensitive processes (CODH, FDH, and hydrogenases), alginate genes, and polymer production and encystment genes. Here, we update information related to genome annotation and phylogeny, and include new material that was not part of that paper because of space limitations.

We do not repeat information from the above paper related to the gene categories listed. The reader interested in learning about these genes should consult the original paper and a follow-up review that contains more information on the genomics of nitrogen fixation, with a focus on *A. vinelandii* (O'Carroll and Dos Santos, 2011).

22.2 MATERIALS AND METHODS

The phylogenetic trees were built using 20 genomes (Table 22.1). For both trees, we used OrthoMCL (Li et al., 2003) to obtain families of orthologous proteins. For the phylogeny of Pseudomonadaceae (Fig. 22.1), OrthoMCL provided 1303 families containing exactly one representative member of each ingroup genome and at most one from the outgroup genome (out of 10,026 families). Each one of these groups was aligned with Muscle (Edgar, 2004), and the noninformative columns were removed by Gblocks (Castresana, 2000). All alignments were concatenated, totaling 399,671 columns, and this final concatenated alignment was used as input to RAxML (Stamatakis, 2006) with the PROTGAMMAWAGF model to build the tree. Bootstrap support values were obtained with 100 replicates. The same method was used to build the tree in Figure 22.2, but in this case we used 6 genomes, 15 families (out of 4989), and 4421 columns. Note that only 80 *Azotobacter chroococcum* protein sequences are available in GenBank.

Gene-sharing information shown in Table 22.3 was obtained from orthoMCL results. AlienHunter (AH) (Vernikos and Parkhill, 2006) was used to identify anomalous regions in the *A. vinelandii* DJ genome. These regions have unusual sequence composition and are considered anomalous if the AH score is above that of an automatically calculated threshold, based on the background sequence composition of the whole genome. For the *A. vinelandii* DJ genome, this threshold was 13.02, and 54 anomalous regions were found. Alignment between chromosomes was carried out using the script Promer from the MUMmer package

Biological Nitrogen Fixation, Volume 1, First Edition. Edited by Frans J. de Bruijn.

Table 22.1 GenBank information on genomes used to create phylogenies

Taxon Name	Taxon ID	Accession Number(s)
Azotobacter vinelandii DJ	322710	NC_012560
Azotobacter chroococcum	355	HD017153.1, HB959743.1, GN125302.1, GM953710.1
Pseudomonas stutzeri A1501	379731	NC_009434
Pseudomonas stutzeri ATCC 17588	96563	NC_015740
Pseudomonas stutzeri CCUG 29243	1196835	NC_018028
Pseudomonas stutzeri DSM 10701	1123519	NC_018177
Pseudomonas stutzeri DSM 4166	996285	NC_017532
Pseudomonas stutzeri RCH2	644801	NC_019936, NC_019937, NC_019938, NC_019939
Pseudomonas putida ND6	231023	NC_005244, NC_017986, NC_018746
Pseudomonas aeruginosa PAO1	208964	NC_002516
Pseudomonas brassicacearum NFM421	994484	NC_015379
Pseudomonas entomophila L48	384676	NC_008027
Pseudomonas fluorescens Pf-5	220664	NC_004129
Pseudomonas fulva 12-X	743720	NC_015556
Pseudomonas mendocina ymp	399739	NC_009439
Pseudomonas putida KT2440	160488	NC_002947
Pseudomonas syringae tomato DC3000	223283	NC_004578, NC_004632, NC_004633
Cellvibrio japonicus Ueda107	498211	NC_010995
Cellvibrio sp. BR	1134474	NZ_AICM00000000
Chromohalobacter salexigens DSM 3043	290398	NC_007963

(Kurtz et al., 2004). Promer performs alignments between translated nucleotide sequences. Promer was not sensitive enough to pick up the alignment between the *iscAnif* gene present in *Teredinibacter turnerae* T7901 genome and that of *A. vinelandii* DJ; this detection required a BLAST search. Clustered regularly interspaced short palindromic repeats (CRISPRs) were found using the CRISPRfinder server (Grissa et al., 2007).

22.3 RESULTS AND DISCUSSION

22.3.1 Genome Features and Phylogeny

The basic genome features of *A. vinelandii* DJ are shown in Table 22.2. They are the same as those reported previously (Setubal et al., 2009), with additional information on group II introns and CRISPRs.

>Setubal et al. (2009) presented a maximum-likelihood phylogeny of Pseudomonadaceae. Figure 22.1 presents a new phylogeny of Pseudomonadaceae, obtained by the same method. This phylogeny has the following novel features compared with the one published previously: it includes two *Cellvibrio* genomes, which are also Pseudomonadaceae but for which in 2009 there were no complete genomes available; five new *Pseudomonas stutzeri* genomes, in addition to *P. stutzeri* A1501; and some newly sequenced *Pseudomonas* species. With the exception of *P. stutzeri*, for all other *Pseudomonas* species we used just one representative per species. This phylogeny is in complete agreement with that from Setubal et al. (2009). The addition of new genomes shows that the *Pseudomonas* species separate into two basic groups: one containing all *P. stutzeri* species and the other containing the remaining *Pseudomonas* species. *A. vinelandii* DJ localizes externally to the *Pseudomonas* species but internally with respect to the two *Cellvibrio* species.

P. stutzeri A1501 is a nitrogen fixer (Yan et al., 2008). Among the five new *P. stutzeri* strains that we included in the phylogenetic inference only one other strain also fixes nitrogen: *P. stutzeri* DSM 4166, an isolate from the rhizosphere of a *Sorghum nutans* cultivar (Yu et al., 2011; see also Chapter 10).

We further investigated the phylogeny of *A. vinelandii* by using the relatively few protein sequences publicly available from *A. chroococcum*. *A. chroococcum* was the first *Azotobacter* species to be described and is commonly found in soils worldwide; it is also a nitrogen fixer (Page, 1987). The resulting phylogeny is shown in Figure 22.2. As expected, *A. chroococcum* groups with *A. vinelandii* DJ, with both species appearing separate from the *Pseudomonas* group.

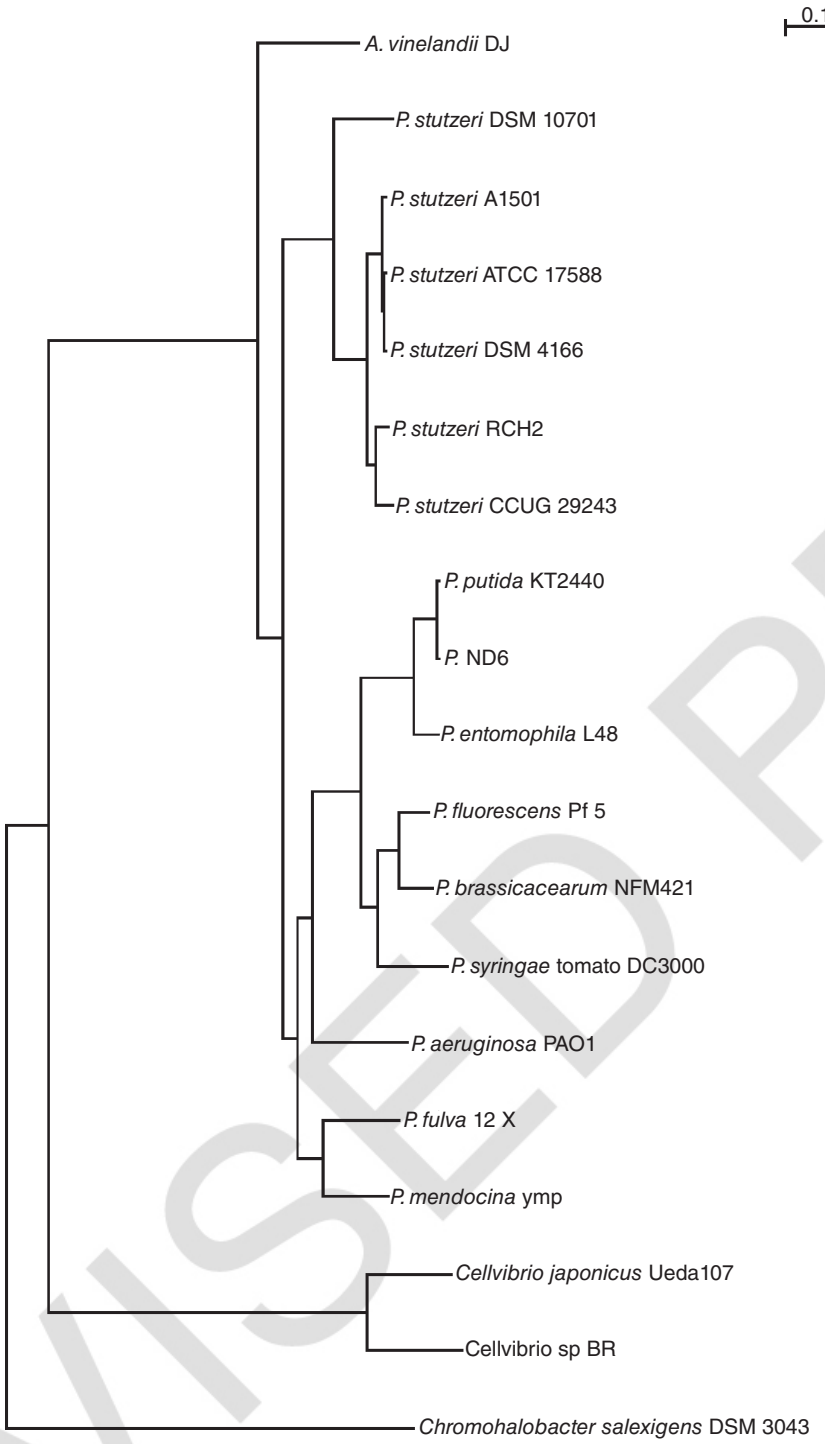

Figure 22.1 A maximum-likelihood phylogenetic tree including representatives from all Pseudomonadaceae species whose genome has been completely sequenced. *Chromohalobacter salexigens* DSM 3043 was used as outgroup. All branches received 100% bootstrap support. The scale shows the number of substitutions per site in branch lengths.

The proximity between *Azotobacter* and *Pseudomonas* shown in these trees justifies the question of whether the genus *Azotobacter* should be reclassified as *Pseudomonas*. Ozen and Ussery (2012) have studied this question, performing extensive comparative analyses. They reached the conclusion that there is "a high similarity between *A. vinelandii* and the *Pseudomonas* genus, suggesting that *Azotobacter* might actually be a *Pseudomonas*." The two trees that we present here show clearly that the two *Azotobacter* species, while related to the sequenced *Pseudomonas*, form a distinct group. Whether or not the *Azotobacter* genus should be reclassified as *Pseudomonas* is a question that we do not address here further.

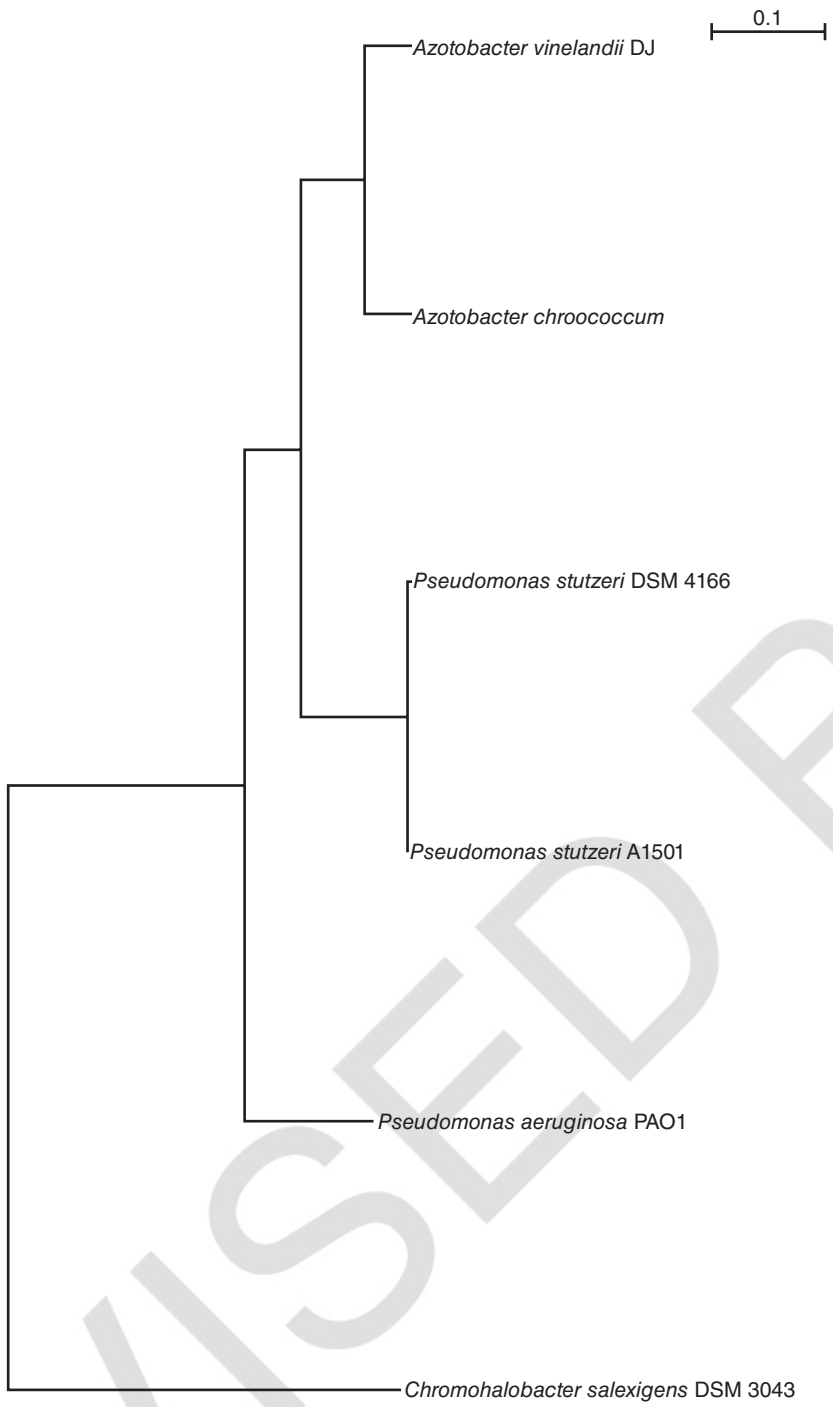

Figure 22.2 A maximum-likelihood phylogenetic tree for some Pseudomonadaceae species including *Azotobacter chroococcum*. All branches received 100% bootstrap support. The scale shows the number of substitutions per site in branch lengths.

22.3.2 Sharing of Nitrogen Fixation-Related Genes

Of all the genomes used to build the phylogeny in Figure 22.1, only *A. vinelandii* DJ, *P. stutzeri* A1501, and *P. stutzeri* DSM 4166 are diazotrophs. We have determined which genes are shared by these three genomes such that none of these genes is also present in the remaining genomes. The results are shown in Table 22.3. With the exception of transposase genes, these three genomes share exactly two regions. One is the major *nif* region (Avin_13060 to Avin_01710); the other is a region (Avin_50900 to Avin_51060) composed of two subregions: the *rnf1* region (Avin_50900 to Avin_50980) and the minor *nif* region

Table 22.2 Basic genome features of *Azotobacter vinelandi* DJ

Feature	*Azotobacter vinelandi* DJ
Size (bp)	5,365,318
%GC	65.7
Total protein-coding genes	5,051
With functional assignment	3,561 (70.5%)
Conserved hypothetical	739 (14.6%)
Hypothetical	751 (14.9%)
Pseudogenes	66
rRNA operons	6
tRNAs	64
Other RNAs	18
Group II introns	8
CRISPRs	3

(Avin_50990 to Avin_51060). Both sets of genes are related to nitrogen fixation (Curatti et al., 2005; O'Carroll and Dos Santos, 2011).

22.3.3 The Origin of Nitrogen Fixation Genes in *A. vinelandii*

There is considerable interest in the origins of nitrogen fixation (Raymond et al., 2004; Kechris et al., 2006; see also Chapters 8, 16, 20). One hypothesis is that nitrogen fixation appeared after the emergence of bacteriochlorophyll biosynthesis and then spread by horizontal gene transfer (HGT) to various microbial lineages (Boyd et al., 2011). The availability of the *A. vinelandii* DJ genome sequence and that of other related organisms presents an opportunity to investigate the issue of HGT for this particular group.

Figure 22.3 shows a graph of anomalous regions in terms of nucleotide composition of the *A. vinelandii* DJ genome. Many of these regions are candidates for having been horizontally transferred; however, none of the nitrogen-fixation genes is contained in any of these regions. This evidence, coupled with the results on gene sharing with the diazotrophic *P. stutzeri* species presented in the previous section, suggests that nitrogen-fixation genes were present in the common ancestor of *A. vinelandii* and *P. stutzeri*, having been lost in the nondiazotrophic *P. stutzeri* species. This means, in turn, that if HGT indeed took place in this lineage, it happened either in that ancestor or earlier.

In order to explore this question further, we have determined which organisms share nitrogen-fixation genes with high similarity to those of *A. vinelandii* and *P. stutzeri*. We carried out this by running BLAST (Altschul et al., 1997) searches using as queries the protein sequences for *A. vinelandii* DJ genes *nifHDKENB* against the nr database from GenBank. These genes have been suggested as a minimum gene set for the computational identification of diazotrophs from whole genome sequences (Dos Santos et al., 2012). Partial results can be seen in Table 22.4. *T. turnerae* T7901 is shown because it gave the best (by bitscore) BLAST hit for all six query sequences among the non-Pseudomonadaceae.

T. turnerae is a marine intracellular endosymbiont Gammaproteobacterium in the order Alteromonadales (NCBI taxonomy). The genome of strain T7901 has been sequenced (Yang et al., 2009), and the similarity of its nitrogen-fixation genes to those of *Azotobacter* and *P. stutzeri* was noted in that publication. Yang et al. (2009) suggested that "the *nif* cluster in *T. turnerae* was acquired via horizontal gene transfer from a *Pseudomonas*-like bacterium." In this context we note that, similar to the *A. vinelandii* results above, none of the *nif* genes in *T. turnerae* T7901 lie in an anomalous region (data not shown). Moreover, an alignment of the *A. vinelandii* DJ genome with that from *T. turnerae* T7901 shows that there is local synteny between the major *nif* regions in both genomes and between the minor *nif* regions in both genomes (Figs. 22.4 and 22.5). In the case of the major *nif* region, the *T. turnerae* chromosome contains an insertion and a deletion with respect to the *A. vinelandii* sequence (Fig. 22.4). The insertion runs from position 1,603,694 to position 1,625,049 in *T. turnerae* genome coordinates and contains genes unrelated to nitrogen fixation, some of which are found scattered in the *A. vinelandii* DJ genome. The deletion occurs between *A. vinelandii* genome coordinates 148,689 and 153,036 (or from Avin_01540 to Avin_01600). This region includes five hypothetical proteins and one ATP-binding cassette (ABC) transporter. This deletion suggests that these genes may not be strictly necessary for nitrogen fixation. In the case of the second region (Fig. 22.5), it essentially corresponds to the second shared region between *A. vinelandii* DJ and the two *P. stutzeri* diazotroph genomes already noted earlier, which includes the minor *nif* region.

These alignments (as well as the global alignment between the two chromosomes – data not shown) are characteristic of genomes that share a common ancestor but which have diverged for a long enough time such that many rearrangements have taken place, but still preserving small syntenic blocks. These results suggest that it is more likely that *T. turnerae*, *A. vinelandii*, and *P. stutzeri* inherited their nitrogen-fixation genes from a common ancestor. If this hypothesis is correct, then we can ask the question: which ancestor, if any, acquired its nitrogen-fixation genes through HGT? Our BLAST search did reveal a few more interesting links, as described next.

We investigated 15 additional non-Pseudomonadaceae species in the BLAST search, selected using the following criteria: only species containing at least five of the six minimum-set genes were selected; and all alignments had to have at least 60% amino acid identity (Table 22.5).

Table 22.3 Genes shared between the *A. vinelandii* DJ, *P. stutzeri* A1501, and *P. stutzeri* DSM 4166 genomes and absent in the other fully sequenced Pseudomonadaceae

Gene (locus tag)	Gene name	Product
Avin_01360	—	Hypothetical protein
Avin_01370	—	Hypothetical protein
Avin_01380	*nifH*	Nitrogenase iron protein
Avin_01390	*nifD*	Nitrogenase molybdenum-iron protein alpha chain: nitrogenase component I, alpha chain
Avin_01400	*nifK*	Nitrogenase molybdenum-iron protein subunit beta
Avin_01410	*nifT*	Nitrogen-fixation protein
Avin_01420	*nifY*	Nitrogenase iron-molybdenum cofactor biosynthesis protein
Avin_01430	—	Hypothetical protein
Avin_01440	*lrv*	Nitrogen-fixing leucine-rich variant repeat 4Fe–4S cluster protein
Avin_01450	*nifE*	Nitrogenase MoFe-cofactor biosynthesis protein
Avin_01470	*nifN*	Nitrogenase molybdenum-cofactor biosynthesis protein
Avin_01480	*nifX*	Nitrogenase MoFe-cofactor biosynthesis protein
Avin_01490	—	Hypothetical protein
Avin_01500	—	Hypothetical protein
Avin_01510	—	Nitrogen fixation (4Fe–4S) ferredoxin-like protein
Avin_01520	*feS1*	Nitrogen fixation (2Fe–2S) ferredoxin (Shethna I protein)
Avin_01530	—	Hypothetical protein
Avin_01540	—	Hypothetical protein
Avin_01550	—	Hypothetical protein
Avin_01560	—	Hypothetical protein
Avin_01570	—	Hypothetical protein
Avin_01610	*iscAnif*	Nitrogen fixation Fe–S cluster assembly protein
Avin_01620	*nifU*	Nitrogen fixation Fe–S cluster scaffold protein
Avin_01630	*nifS*	Nitrogen fixation cysteine desulfurase
Avin_01650	*cysE1*	Nitrogen fixation serine O-acetyltransferase
Avin_01660	—	Hypothetical protein
Avin_01670	*nifW*	Nitrogen fixation protein
Avin_01680	*nifZ*	Nitrogen fixation protein
Avin_01690	*nifM*	Nitrogen fixation cis-trans peptidyl prolyl isomerase
Avin_01700	*clpX*	ATP-dependent protease ATP-binding subunit
Avin_01710	*nifF*	Flavodoxin
Avin_09770	—	Transposase IS3/IS911 (5 copies in Avin)
Avin_50900	—	Nitrogen fixation-like protein
Avin_50910	*nafY*	Nitrogen fixation-like protein subunit gamma
Avin_50920	*rnfH*	RnfABCDGE type electron transport complex subunit H
Avin_50930	*rnfE1*	RnfABCDGE type electron transport complex subunit E
Avin_50940	*rnfG1*	RnfABCDGE type electron transport complex subunit G
Avin_50950	*rnfD1*	RnfABCDGE type electron transport complex subunit D
Avin_50960	*rnfC1*	RnfABCDGE type electron transport complex subunit C
Avin_50970	*rnfB1*	RnfABCDGE type electron transport complex subunit B
Avin_50980	*rnfA1*	RnfABCDGE type electron transport complex subunit A
Avin_50990	*nifL*	Nitrogen fixation regulatory protein
Avin_51010	*nifB*	Nitrogenase cofactor biosynthesis protein
Avin_51020	—	Ferredoxin protein
Avin_51030	—	Nitrogenase-associated protein
Avin_51040	*nifO*	Nitrogen fixation cofactor assembly protein
Avin_51050	*nifQ*	Rhodanese/sulfurtransferase-like protein
Avin_51060	—	Glutaredoxin-like protein

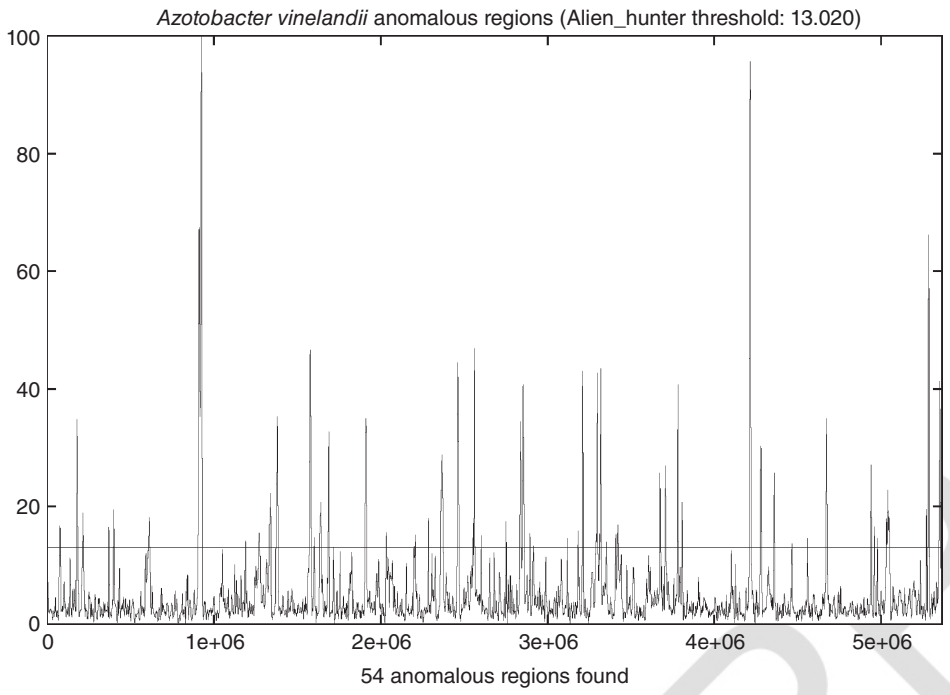

Figure 22.3 Anomalous regions in the genome of *A. vinelandii* DJ genome. The *x*-axis represents genome coordinates and the *y*-axis represents AlienHunter scores, calculated for each 5 kbp-length window with an offset of 2.5 kbp.

Of these 16 species (including *T. turnerae*), 12 are Gammaproteobacteria (with seven species in the order Chromatiales), and four are Betaproteobacteria.

Our interpretation of these results is that the presence of nitrogen-fixation genes in *A. vinelandii* is the result of vertical inheritance up to an unidentified Gammaproteobacterium ancestor. Because that hypothesized ancestor is also the ancestor of many nondiazotrophs, such a hypothesis requires that nitrogen-fixation genes were lost in many descendant lineages of that ancestor. The above-mentioned evidence suggests that HGT may have happened between that Gammaproteobacterium ancestor and a Betaproteobacterium ancestor, given that the connection between these two groups was one of the "highways" for HGT proposed by Beiko et al. (2005). Additional research is required to determine whether nitrogen-fixation genes were indeed exchanged in ancient HGT events between Betaproteobacteria and Gammaproteobacteria.

Table 22.4 Similarity of *A. vinelandii* DJ protein sequences given by percent identity with sequences from the following organisms: *P. stutzeri* A1501, *P. stutzeri* DSM 4166, and *Teredinibacter turnerae* T7901

nif Gene	A1501	DSM 4166	*Teredinibacter turnerae* T7901
H	92	92*	90
D	91	91	86
K	91	90	82
E	89	88	83
N	81	80	67
B	86*	86	89

The results were obtained with BLAST. For those marked with an asterisk, it was necessary to use tBLASTn to obtain the correct result because of annotation problems. All alignments covered 100% or nearly 100% of both query and subject sequences.

22.3.4 Other Features of the *A. vinelandii* DJ Genome

The *A. vinelandii* DJ genome has the remarkable property that it contains eight group II introns (Fig. 22.6), more than any other bacterial genome (Dai et al., 2003). (The database on which this statement is based was last updated in 2008.) Group II introns are a class of RNAs that can perform a self-splicing reaction. Bacterial group II introns almost always encode reverse transcriptase ORFs (Open Reading Frames) and are active mobile elements (Dai and Zimmerly, 2002). All of the identified group II introns in *A. vinelandii* DJ encode reverse transcriptase ORFs. These introns have been carefully annotated and are part of the *A. vinelandii* DJ GenBank record.

Another notable feature of the *A. vinelandii* DJ genome is that it contains three sets of CRISPRs. CRISPRs are thought to constitute a kind of RNA-interference-based immune system for prokaryotes (Makarova et al., 2006). We report them here primarily because they are not currently annotated in the *A. vinelandii* DJ GenBank record. Two of the sets display the expected structure for CRISPRs: from position 3261022 to position 3263057, with 30 spacers, and from position 3265050 to position 3266352, with 19 spacers. A third set has "questionable structure" (Grissa et al., 2007),

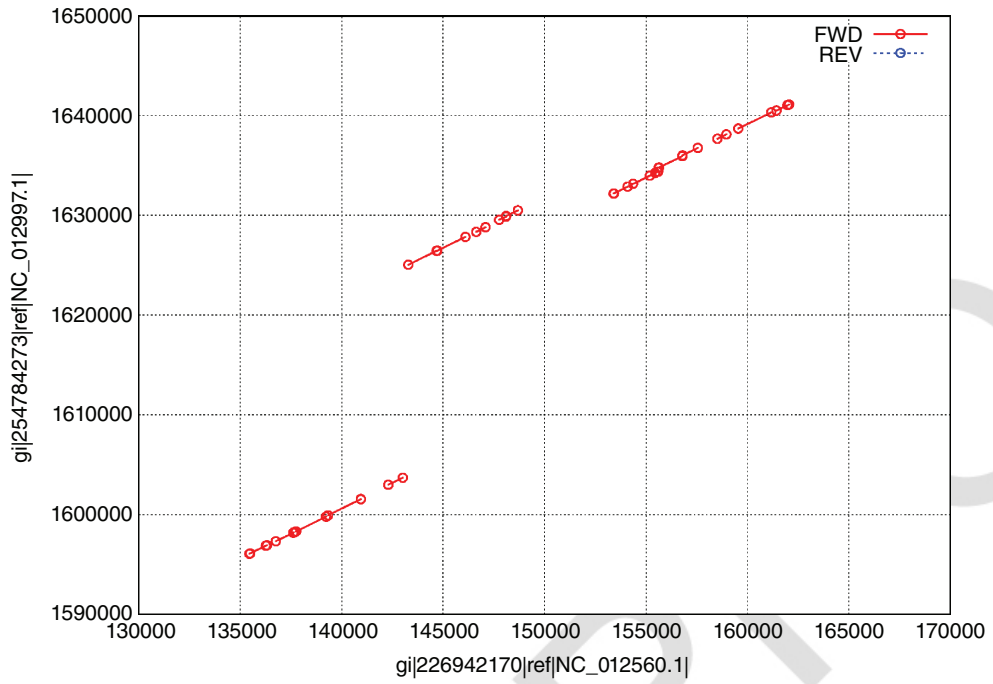

Figure 22.4 Alignment between the chromosome sequence of *A. vinelandii* DJ (*x*-axis) and the chromosome sequence of *T. turnerae* T7901 (*y*-axis). Only the regions delimited by the positions shown in both axes are shown. The region shared by both corresponds to the major *nif* region, with the modifications noted in the text.

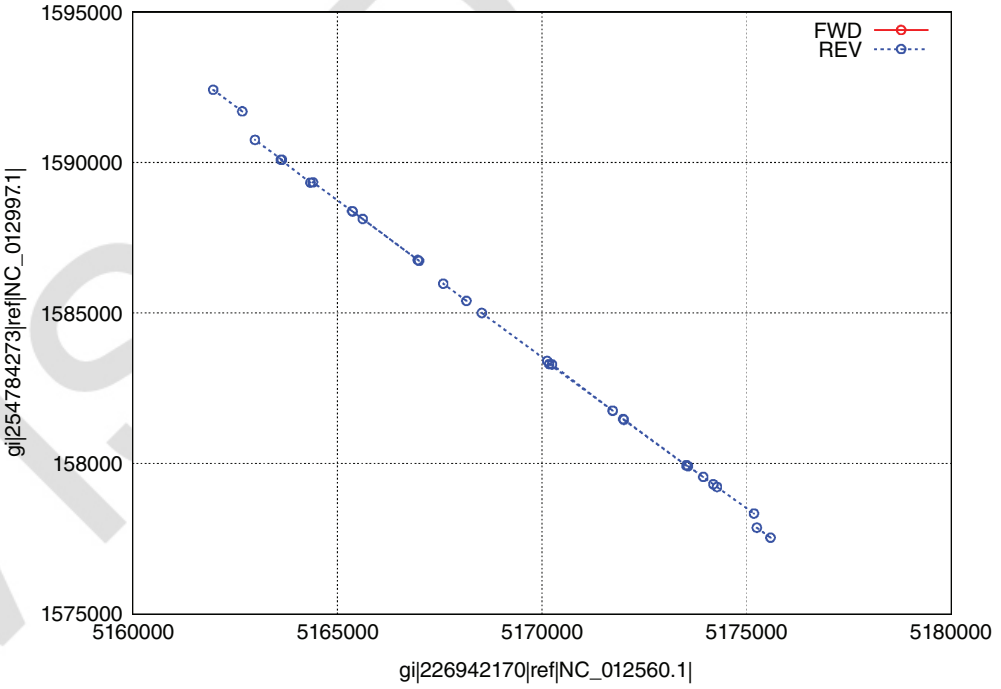

Figure 22.5 Alignment between the chromosome sequence of *A. vinelandii* DJ (*x*-axis) and the chromosome sequence of *T. turnerae* T7901 (*y*-axis). Only the regions delimited by the positions shown in both axes are shown. The region shared by both (which corresponds to Avin_50910 to Avin_51060) includes the minor *nif* region.

Table 22.5 BLAST results Using as queries genes *nifB*, *nifD*, *nifE*, *nifH*, *nifK*, and *nifN* from *A. vinelandii* DJ

	B	D	E	H	K	N
Allochromatium vinosum DSM 180 (gamma)	73/753	84/876	76/756	89/533	79/885	58/542
Azoarcus sp. BH72 (beta)*	72/734	81/861	77/769	87/517	77/874	60/543
Beggiatoa alba B18LD (gamma)	76/780	83/878	80/806	87/525	79/899	63/580
Candidatus Accumulibacter phosphatis clade IIA str. UW-1 (beta)*	72/727	80/842	78/731	86/523	77/872	60/538
Dechloromonas aromatica RCB (beta)*	73/719	81/845	79/765	87/496	75/843	61/560
Methylobacter tundripaludum SV96 (gamma)	75/802	81/872	79/809	86/511	77/878	57/530
Methylomonas methanica MC09 (gamma)	76/783	82/871	81/807	86/508	79/903	59/551
Sideroxydans lithotrophicus ES-1 (beta)*	73/757	82/862	78/778	88/534	78/887	62/569
Teredinibacter turnerae T7901 (gamma)	82/841	86/910	83/827	90/539	82/939	67/653
Thiocapsa marina 5811 (gamma)	74/750	83/882	77/771	89/537	78/887	60/525
Thiocystis violascens DSM 198 (gamma)	75/779	84/892	78/753	89/531	79/892	59/553
Thioflavicoccus mobilis 8321 (gamma)	70/716	82/861	80/779	88/523	76/872	63/555
Thiorhodococcus drewsii AZ1 (gamma)	75/768	84/872	77/791	88/532	80/900	58/537
Thiorhodospira sibirica ATCC 700588 (gamma)	Not found	83/859	77/759	88/532	79/905	60/499
Thiorhodovibrio sp. 970 (gamma)	60/546	81/868	75/756	89/535	76/869	59/547
Thiothrix nivea DSM 5205 (gamma)	76/762	82/874	79/801	89/533	77/872	62/540

The first column contains the organism found (with proteobacterial class in parentheses), and the cells in that row show percent identity/bit score for each hit. For criteria in selecting hits, see the text. Asterisks indicate Betaproteobacteria. BLAST search was done on March 25, 2013.

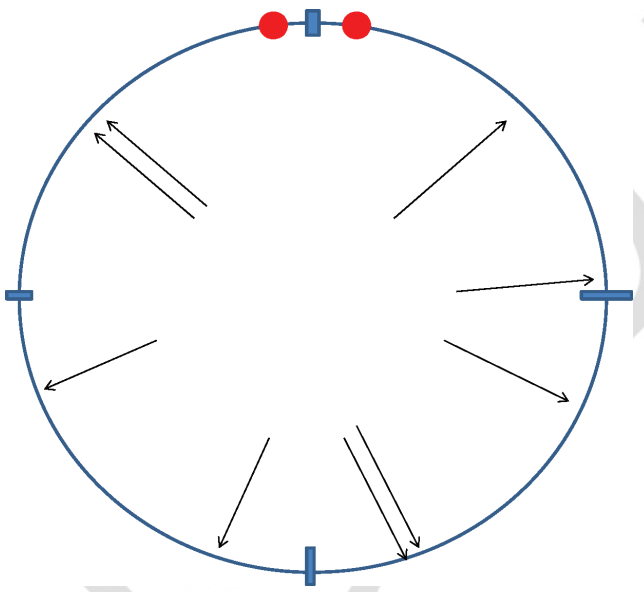

Figure 22.6 Approximate positions of group II introns in the chromosome sequence of *A. vinelandii* DJ (arrows). The two red dots represent the location of the minor (left) and major (right) *nif* regions. At the top is position 1.

and is located from position 3171496 to position 3171619, with two spacers. In addition, the genome contains two cas (CRISPR-associated) groups of protein-coding genes: Avin_17170 to Avin_17240 (spanning region 1,699,773 through 1,707,301 bp) and Avin_31570 to Avin_31630 (spanning region 3,266,532–3,273,695 bp). The second cas operon is thus immediately downstream from the second CRISPR set noted earlier. These data are strong evidence that *A. vinelandii* DJ has a functioning CRISPR-Cas system (CASS).

It has been proposed that a CASS can work as a defensive mechanism against bacteriophage invasion (Makarova et al., 2006). The *A. vinelandii* DJ genome has 40 protein-coding genes annotated with the word "phage." Several of these are scattered throughout the genome and do not form a complete prophage. But 21 of those genes can be found in the region defined by genes Avin_37340 to Avin_37610 (approximately 22 kbp). Part of this region (~5.5 kbp) does match (by MEGABLAST (Altschul et al., 1997)) part of *Pseudomonas aeruginosa* phage phiCTX (Nakayama et al., 1999). But this may be an ancient insertion (the region was not picked up by AH as anomalous), and the whole region may no longer contain a functional prophage. It is a matter however that deserves further investigation.

22.4 CONCLUSION

A. vinelandii is an important model organism for the study of nitrogen fixation and iron–sulfur clusters. The publication of the *A. vinelandii* DJ genome has enabled or facilitated various kinds of studies in these areas. Notable among them was the first transcriptome analysis of *A. vinelandii*, by Hamilton et al. (2011), who studied expression levels of the three nitrogen-fixation systems Nif, Anf, and Vnf that exist in *A. vinelandii* (see also Chapter 9).

Here, results concerning the genome of *A. vinelandii* DJ have been presented, updated, and expanded. Among these, the possible origin of its nitrogen-fixing genes was investigated, which led to the hypothesis that they were inherited

from a common ancestor of *A. vinelandii*, *P. stutzeri*, and *T. turnerae*. Information on the presence of group II introns and the CASS system has also been provided. It is hoped that the results presented here will further facilitate research that depends on the *A. vinelandii* DJ genome.

ACKNOWLEDGMENTS

We would like to thank Ray Dixon for a critical reading of an earlier version of the manuscript and the Conselho Nacional de Desenvolvimento Científico e Tecnológico and Fundect-MS (Grant TO 0096/12) for partial funding.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17): 3389–3402.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102(40): 14332–14337.

Boyd ES, Hamilton TL, Peters JW. 2011. An alternative path for the evolution of biological nitrogen fixation. Front. Microbiol. 2: 205.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17(4): 540–552.

Curatti L, Brown CS, Ludden PW, Rubio LM. 2005. Genes required for rapid expression of nitrogenase activity in *Azotobacter vinelandii*. Proc. Natl. Acad. Sci. U. S. A. 102(18): 6291–6296.

Dai L, Toor N, Olson R, Keeping A, Zimmerly S. 2003. Database for mobile group II introns. Nucleic Acids Res. 31(1): 424–426.

Dai L, Zimmerly S. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic Acids Res. 30(5): 1091–1102.

Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 13: 162.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792–1797.

Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 35(Web Server issue): W52–W57.

Hamilton TL, Ludwig M, Dixon R, Boyd ES, Dos Santos PC, Setubal JC, et al. 2011. Transcriptional profiling of nitrogen fixation in *Azotobacter vinelandii*. J. Bacteriol. 193(17): 4477–4486.

Kechris KJ, Lin JC, Bickel PJ, Glazer AN. 2006. Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. Proc. Natl. Acad. Sci. U. S. A. 103(25): 9584–9589.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5(2): R12.

Li L, Stoeckert, Jr., CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9): 2178–2189.

Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol. Direct 1: 7.

Nakayama K, Kanaya S, Ohnishi M, Terawaki Y, Hayashi T. 1999. The complete nucleotide sequence of phi CTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa*: implications for phage evolution and horizontal gene transfer via bacteriophages. Mol. Microbiol. 31(2): 399–419.

O'Carroll IP, Dos Santos PC. 2011. Genomic analysis of nitrogen fixation. Methods Mol. Biol. 766: 49–65.

Ozen AI, Ussery DW. 2012. Defining the *Pseudomonas* genus: where do we draw the line with *Azotobacter*? Microb. Ecol. 63(2): 239–248.

Page WJ. 1987. Iron-dependent production of hydroxamate by sodium-dependent *Azotobacter chroococcum*. Appl. Environ. Microbiol. 53(7): 1418–1424.

Raymond J, Siefert JL, Staples CR, Blankenship RE. 2004. The natural history of nitrogen fixation. Mol. Biol. Evol. 21(3): 541–554.

Setubal JC, Dos Santos PC, Goldman BS, Ertesvag, H, Espin G, Rubio LM, et al. 2009. Genome sequence of *Azotobacter vinelandii*, an obligate aerobe specialized to support diverse anaerobic metabolic processes. J. Bacteriol. 191(14): 4534–4545.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21): 2688–2690.

Vernikos GS, Parkhill J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics 22(18): 2196–2203.

Yan Y, Yang J, Dou Y, Chen M, Ping S, Peng J, et al. 2008. Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated *Pseudomonas stutzeri* A1501. Proc. Natl. Acad. Sci. U. S. A. 105(21): 7564–7569.

Yang JC, Madupu R, Durkin AS, Ekborg, NA, Pedamallu CS, Hostetler JB, et al. 2009. The complete genome of *Teredinibacter turnerae* T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms). PLoS One 4(7): e6085.

Yu H, Yuan M, Lu W, Yang J, Dai S, Li Q, et al. 2011. Complete genome sequence of the nitrogen-fixing and rhizosphere-associated bacterium *Pseudomonas stutzeri* strain DSM 4166. J. Bacteriol. 193(13): 3422–3423.