The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability

Brazilian National Genome Project Consortium*

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved July 7, 2003 (received for review April 11, 2003)

Chromobacterium violaceum is one of millions of species of free-living microorganisms that populate the soil and water in the extant areas of tropical biodiversity around the world. Its complete genome sequence reveals (*i*) extensive alternative pathways for energy generation, (*ii*) ≈500 ORFs for transport-related proteins, (*iii*) complex and extensive systems for stress adaptation and motility, and (*iv*) widespread utilization of quorum sensing for control of inducible systems, all of which underpin the versatility and adaptability of the organism. The genome also contains extensive but incomplete arrays of ORFs coding for proteins associated with mammalian pathogenicity, possibly involved in the occasional but often fatal cases of human *C. violaceum* infection. There is, in addition, a series of previously unknown but important enzymes and secondary metabolites including paraquat-inducible proteins, drug and heavy-metal-resistance proteins, multiple chitinases, and proteins for the detoxification of xenobiotics that may have biotechnological applications.

The genomes of soil- and water-borne free-living bacteria have received relatively little attention thus far in comparison to pathogenic and extremophilic organisms, yet they provide fundamental insights into environmental adaptation strategies and represent a rich source of genes with biotechnological potential and medical utility. A particularly interesting organism of this kind is *Chromobacterium violaceum*, a Gram-negative β-proteobacterium first described at the end of the 19th century (1), which dominates a variety of ecosystems in tropical and subtropical regions. This bacterium has been found to be highly abundant in the water and borders of the Negro river, a major component of the Brazilian Amazon (2) and as a result has been studied in Brazil over the last three decades. These, in general, have focused on the most notable product of the bacterium, the violacein pigment, which has already been introduced as a therapeutic compound for dermatological purposes (3). Violacein also exhibits antimicrobial activity against the important tropical pathogens *Mycobacterium tuberculosis* (4), *Trypanosoma cruzi* (5), and *Leishmania* sp. (6) and is reported to have other bactericidal (2, 7–10), antiviral (11), and anticancer (12, 13) activities.

Some other aspects of the biotechnological potential of *C. violaceum* have also begun to be explored, including the synthesis of poly(3-hydroxyvaleric acid) homopolyester and other short-chain polyhydroxyalkanoates, which might represent alternatives to plastics derived from petrochemicals (14, 15), the hydrolysis of plastic films (16), and the solubilization of gold through a mercury-free process, thereby avoiding environmental contamination (17, 18). These studies, however, have been based on knowledge of only a tiny fraction of the genetic constitution of the organism. In addition, the more basic issues of the mechanisms and strategies underlying the adaptability of *C. violaceum*, including its observed but infrequent infection of humans, have not been deeply investigated at the molecular and genetic levels.

To begin to rectify the paucity of our basic knowledge of this remarkable organism we sequenced and annotated the complete genome of *C. violaceum* type strain ATCC 12472. This has revealed a detailed portrait of the molecular complexity required for the organism's versatility as well as an extended compendium of ORFs that significantly increase the biotechnological potential of the bacterium.

Materials and Methods

The sequencing and analysis of the *C. violaceum* genome were entirely executed by the Brazilian National Genome Sequencing Consortium comprising 25 sequencing laboratories, 1 bioinformatics center, and 3 coordination laboratories distributed throughout Brazil.

*Brazilian National Genome Project Consortium: Ana Tereza Ribeiro de Vasconcelos[a], Darcy F. de Almeida[b], Mariangela Hungria[c], Claudia Teixeira Guimarães[d], Regina Vasconcellos Antônio[e], Francisca Cunha Almeida[f], Luiz G. P. de Almeida[a], Rosana de Almeida[g], José Antonio Alves-Gomes[h], Elizabeth Mazoni Andrade[i], Julia Araripe[j], Magnólia Fernandes Florêncio de Araújo[k], Spartaco Astolfi-Filho[l], Vasco Azevedo[i], Alessandra Jorge Baptista[m], Luiz Artur Mendes Bataus[n], Jacqueline da Silva Batista[h], André Beló[p], Cássio van den Berg[p], Maurício Bogo[r], Sandro Bonatto[r], Juliano Bordignon[s], Marcelo Macedo Brigido[m], Cristiana Alves Brito[t], Marcelo Brocchi[g], Helio Almeida Burity[u], Anamaria Aranha Camargo[v], Divina das Dores de Paula Cardoso[w], Newton Portilho Carneiro[d], Dirce Maria Carraro[v], Cláudia Márcia Benedetto Carvalho[i], Júlio Cézar de Mattos Cascardo[x], Benildo Sousa Cavada[y], Ligia Maria O. Chueire[c], Tânia Beatriz Creczynski-Pasa[z], Nivaldo Costa da Cunha-Junior[c], Nelson Fagundes[r], Clarissa Lima Falcão[o], Fabiana Fantinatti[aa], Izeni Pires Farias[l], Maria Sueli Soares Felipe[m], Lilian Pereira Ferrari[p], Jesus Aparecido Ferro[bb], Maria Inês Tiraboschi Ferro[bb], Gloria Regina Franco[t], Nara Suzy Aguiar de Freitas[cc], Luiz Roberto Furlan[dd], Ricardo Tostes Gazzinelli[t], Eliane Aparecida Gomes[d], Pablo Rodrigues Gonçalves[f], Thalles Barbosa Grangeiro[y], Dario Grattapaglia[o], Edmundo Carlos Grisard[s], Ebert Seixas Hanna[g], Sílvia Neto Jardim[d], Jomar Laurino[r], Lélia Cristina Tenório Leoi[o], Lucymara Fassarella Agnez Lima[ee], Maria de Fatima Loureiro[c], Maria do Carmo Catanho Pereira de Lyra[cc], Humberto Maciel França Madeira[p], Gilson Paulo Manfio[aa], Andrea Queiroz Maranhão[m], Wellington Santos Martins[o], Sônia Marli Zingaretti di Mauro[bb], Silvia Regina Batistuzzo de Medeiros[ee], Rosely de Vasconcellos Meissner[k], Miguel Angelo Martins Moreira[f], Fabrícia Ferreira do Nascimento[f], Marisa Fabiana Nicolás[c], Jaquelline Germano Oliveira[t], Sergio Costa Oliveira[t], Roger Ferreira Cury Paixão[a], Juliana Alves Parente[n], Fabio de Oliveira Pedrosa[ff], Sergio Danilo Junho Pena[t], José Odair Pereira[gg], Maristela Pereira[n], Luciana Santos Rodrigues Costa Pinto[x], Luciano da Silva Pinto[y], Jorge Ivan Rebelo Porto[h], Deise Porto Potrich[hh], Cicero Eduardo Ramalho-Neto[ii], Alessandra Maria Moreira Reis[o], Liu Um Rigo[ff], Edson Rondinelli[j], Elen Bethleen Pedraça do Santos[l], Fabrício R. Santos[i], Maria Paula Cruz Schneider[kk], Hector N. Seuanez[f,q], Ana Maria Rodrigues Silva[m], Artur Luiz da Costa da Silva[kk], Denise Wanderlei Silva[ii], Rosane Silva[i], Isabella de Carmo Simões[m], Daniel Simon[r], Célia Maria de Almeida Soares[n], Renata de Bastos Ascenço Soares[n], Emanuel Maltempi Souza[ff], Kelly Rose Lobo de Souza[f], Rangel Celso Souza[a], Maria Berenice Reynaud Steffens[ff], Mário Steindel[s], Santuza Ribeiro Teixeira[t], Turan Urmenyi[j], André Vettore[v], Roseli Wassem[ff], Arnaldo Zaha[hh], and Andrew John George Simpson[v,ll].

[kk]Department of Genetics, Federal University of Pará, Campus Universitário Guamá, Caixa Postal 8607, CEP 66.075-970, Belém, PA, Brazil; [a]Labinfo, Laboratorio Nacional de Computação Científica/Ministério da Ciência e Tecnologia, Rua Getúlio Vargas 333, CEP 25651-071, Petrópolis, RJ, Brazil; [b]Institute of Biophysics Carlos Chagas Filho, Federal University of Rio de Janeiro, Cidade Universitaria, CEP 21941-590, Rio de Janeiro, RJ, Brazil; [bb]Department of Technology, Universidade Estadual Paulista, CEP 14884-900, Jaboticabal, SP, Brazil; [dd]Department of Animal Nutrition, Universidade Estadual Paulista, CEP 18610-000, Botucatu, SP, Brazil; [m]Department of Cellular Biology, University of Brasilia, Institute of Biological Sciences, CEP 70910-900, Brasília, DF, Brazil; [ff]Department of Biochemistry and Molecular Biology, Federal University of Paraná, Centro Politécnico, Caixa Postal 19046, CEP 81531-990, Curitiba, PR, Brazil; [hh]Department of Molecular Biology and Biotechnology, Centro de Biotecnologia, Federal University of Rio Grande do Sul, Avenida Bento Gonçalves, 9500, Caixa Postal 15.005, CEP 91.501-970, Porto Alegre, RS, Brazil; [i]Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Avenida Antônio Carlos, 6627, Caixa Postal 486, CEP 31270-010, Belo Horizonte, MG, Brazil; [ii]Department of Molecular Genetics, Genomics, and Proteomics, Federal University of

Sequencing and Assembly. The *C. violaceum* type strain ATCC 12472 was used as DNA source for the construction of cosmid libraries in Lawrist 4 and short insert libraries in pUC18 as described elsewhere (19, 20). Template preparation and DNA sequencing reactions were performed by using standard protocols. The latter used DYEnamic ET dye terminator cycle sequencing (MegaBACE) and the MegaBACE 1000 capillary sequencer (Amersham Pharmacia Biotech). Approximately 80,000 reads with PHRED scores >20 were generated from both ends of plasmid clones ranging from 2.0 to 4.0 kb, providing a 13-fold genome coverage. These sequences were assembled by using PHRED/PHRAP/CONSED (www.phrap.org). Both ends of 3,350 cosmid clones with an average 40-kb insert size were also sequenced, providing a validation check of the final assembly. Sequencing gaps were closed by using the information generated by autofinisher. A new strategy, PCR-assisted contig extension (21), was also used for physical gap closure.

Genome Annotation. Annotation was carried out by using the system for automated bacterial integrated annotation (unpublished data), developed to integrate public domain and purpose-built software for the automated identification of genome landmarks including

Alagoas, Campus Delza Gitai km. 85 BR 104 Norte, CEP 57100-000, Rio Largo, AL, Brazil; [o]Department of Genomic Sciences and Biotechnology, Catholic University of Brasília, 916 Norte CEP 70.790-160, Brasília, DF, Brazil; [aa]Centro Pluridisciplinar de Pesquisas Químicas, Biológicas e Agrícolas (CPQBA), Divisão de Recursos Microbianos (DRM), Campinas State University-UNICAMP, Caixa Postal CP 6171, CEP 13083-970, Campinas, SP, Brazil; [r]Faculty of Biosciences, Center of Genomic and Molecular Biology, Rio Grande do sul Pontifical Catholic University, Avenida Ipiranga 6681-Prédio 12C, CEP 90619-900, Porto Alegre, RS, Brazil; [s]Department of Microbiology and Parasitology, Federal University of Santa Catarina, Ciências Biológicas, Campus Universitário, Trindade, Caixa Postal 476, CEP 88040-900, Florianópolis, SC, Brazil; [d]Department of Applied Biology, Embrapa Milho e Sorgo, Caixa Postal 151, CEP 35701-970, Sete Lagoas, MG, Brazil; [x]Department of Biological Sciences, State University of Santa Cruz, Ilheus-Itabuna Road, km. 16, CEP 45650-000, Ilheus, BA, Brazil; [n]Department of Biochemistry, Institute of Biological Science Institution, Federal University of Goias, Campus Samambaia, CEP 74001-970, Goiania, GO, Brazil; [h]Coordination of Research in Aquatic Biology, Instituto Nacional de Pesquisas da Amazônia, Avenida André Araujo, 2936, Caixa Postal 480, CEP 69060-001, Manaus, AM, Brazil; [ee]Department of Cellular Biology and Genetics, Center of Biosciences, Federal University of Rio Grande do Norte, Campus Universitário, Lagoa Nova, CEP 59076-700, Natal, RN, Brazil; [l]Biology Department, Amazonas Federal University, Avenida Rodrigo Otávio Jordão Ramos 3000, CEP 69077-000, Manaus, AM, Brazil; [y]Department of Biochemistry and Molecular Biology, Federal University of Ceara, Campus do Pici, s/n bl. 907 CP 6033, CEP 6041-970, Fortaleza, CE, Brazil; [f]Genetics Division-Diretoria de Pesquisa, Instituto Nacional de Câncer, Rua André Cavalcanti 37, CEP 20231-050, Rio de Janeiro, RJ, Brazil; [g]Department of Cellular and Molecular Biology, School of Medicine at Ribeirao Preto, University of Sao Paulo, Avenida Bandeirantes, 3900, CEP 14049-900, Ribeirao Preto, SP, Brazil; [c]Laboratory of Biotechnology of Soils, Embrapa Soja, Caixa Postal 231, CEP 86-001970, Londrina, PR, Brazil; [v]Laboratory of Genetics, Ludwig Institute for Cancer Research, Rua Professor Antonio Prudente, 109/4°. andar, CEP 01509-010, São Paulo, SP, Brazil; [cc]Department of Biology, Rural Federal University of Pernambuco, Rua Dom Manuel de Medeiros, CEP 52171-930, Dois Irmãos, Recife, PE, Brazil; [u]EMBRAPA/Empresa Pernambucana de Pesquisa Agropecuária, Recife, PE, Brazil; [p]Center for Agricultural and Environmental Sciences, Paraná Pontifical Catholic University, Rod. BR-376, km. 14, CEP 83010-500, São José dos Pinhais, PR, Brazil; [q]Department of Genetics, Federal University of Rio de Janeiro, Caixa Postal 68011, CEP 21944-970, Rio de Janeiro, RJ, Brazil; [w]Department of Microbiology, Immunology, Parasitology, and Pathology, Institute of Tropical Pathology and Public Health, Federal University of Goiás, Rua Delenda Rezende de Melo, Setor Universitário, CEP 74605-050, Goiânia, GO, Brazil; [z]Department of Pharmaceutical Sciences, Federal University of Santa Catarina, Campus Universitário-Trindade, Caixa Postal 476, CEP 88040-900, Florianópolis, SC, Brazil; [k]Department of Microbiology and Parasitology, Center of Biosciences, Federal University of Rio Grande do Norte, Campus Universitário, Lagoa Nova, CEP 59076-700, Natal, RN, Brazil; [t]Department of Biochemistry and Immunology, Institute of Biological Sciences, Federal University of Minas Gerais, Avenida Antônio Carlos, 6627, Caixa Postal 486, CEP 31270-901, Belo Horizonte, MG, Brazil; [e]Department of Biochemistry, Federal University of Santa Catarina, Campus Universitário, Trindade, Caixa Postal 470, CEP 88040-900, Florianópolis, SC, Brazil; [gg]Department of Fundamental Science and Agricola Development, Amazonas Federal University, Avenida Rodrigo Otávio Jordão Ramos 3000, CEP 69077-000, Manaus, AM, Brazil; [j]Department of Molecular and Structural Biology, Carlos Chagas Filho Biophysics Institute, Federal University of Rio de Janeiro, Bl. G, Centro de Ciencias da Saude, Cidade Universitaria, CEP 21949-900, Rio de Janeiro, RJ, Brazil; and [jj]Department of Internal Medicine, School of Medicine, Carlos Chagas Filho Biophysics Institute, Federal University of Rio de Janeiro, Cidade Universitaria, CEP 21.949-900, Rio de Janeiro, RJ, Brazil.

[ll]To whom correspondence should be addressed at: Ludwig Institute for Cancer Research, 605 Third Avenue, New York, NY 10158. E-mail: asimpson@licr.org.

Table 1. General features of the *C. violaceum* genome

Length, bp	4,751,080
G + C content	64.83%
Total no. of ORFs	4,431
Percentage of genome constituting coding regions	89%
Average ORF length, bp	954
No. of known proteins	2,717
No. of conserved hypothetical proteins	958
No. of hypothetical proteins	756
rRNAs	8 × (16S-23S-5S)
tRNAs	98

tRNA and rRNA genes, repetitive elements, and ORFs likely to encode proteins. For putative functional attribution, BLAST programs (www.ncbi.nlm.nih.gov) were used to search for similarity in the main sequence databases. These results were instrumental in identifying metabolic pathways based on the *Kyoto Encyclopedia of Genes and Genomes* (22). For comparison of protein sequences between species, we used COG (23), INTERPRO (24), PRINTS (www.bioinf.man.ac.uk/dbbrowser/PRINTS), PSORT (25), and TCDB (http://tcdb.ucsd.edu/tcdb). Noncoding regions were annotated by using software that seeks ribosomal binding sites for the identification of promoters and operators. Paralogous gene families were defined by using a cutoff E value of 10^{-5} with at least 60% query coverage and 50% identity.

Results and Discussion

General Features of the Genome. The complete genome of the *C. violaceum* consists of a single circular chromosome of 4,751,080 bp with an average G+C content of 64.83% (see Table 1 and supplementary information at www.brgene.lncc.br/cviolaceum; GenBank accession no. AE016825). There are 4,431 uniformly distributed predicted protein coding ORFs that cover 89% of the genome and have an average length of 954 bp. Of these, 2,717 (61.3%) could be assigned putative functions, whereas 958 (21.6%) were identified as conserved hypothetical proteins. The remaining 756 (17.1%) were designated hypothetical proteins. Of the conserved hypothetical ORFs, 499 have protein motifs contained within both INTERPRO and COG, whereas 242 have motifs contained in either one or the other. Among the hypothetical ORFs, 68 have motifs contained in both and 135 in only one of the two databases. Of the 131 paralogous families, 111 (84.7%) contain two members, but some contain as many as six ORFs. The functions of approximately one-third of the families are related to transport, and approximately one-fourth have unknown functions (see supplementary information at www.brgene.lncc.br/cviolaceum). There are 98 tRNA genes representing all 20 amino acids and 8 rRNA operons that are identical in their coding region, although 6 contain a 100-bp insert in the spacer region. The likely origin of replication is identifiable based on G+C skew and the positions of *dnaA*, *dnaN*, and *gyrA* (26).

Comparison with Other Sequenced Genomes. Comparison of the *C. violaceum* ORFs with those of other organisms reveals that 17.4% have closest similarity to ORFs of *Ralstonia solanacearum* (27), a soil-borne phytopathogen (27); 9.75% to ORFs of *Neisseria meningitidis* serogroup A, the causal agent of a serious human disease (28); and 9.61% to ORFs of *Pseudomonas aeruginosa*, a free-living bacterium causing opportunistic infections in humans (29) (see supplementary information at www.brgene.lncc.br/cviolaceum). The ORFs with highest similarity to *R. solanacearum* are mostly from COG categories N–Q (cell motility, posttranslational modification, inorganic ion transport, and secondary metabolite biosynthesis, respectively) and thus are directly related to the bacterium's interactions with the environment. Approximately half (50.1%) of these ORFs with highest similarity with *R. solanacearum* are absent from *N. meningitides*. This suggests that they may be restricted to free-living organisms. Thus, environmental adaptation is to some

MICROBIOLOGY

Table 2. Comparative distribution of ORF function among selected free-living organisms

	cv*	bs*	ec*	dr*	tm*	pa*	sc*	xcc*	pp*
cog categories									
C, energy production and conversion	204	168	275	110	109	305	345	182	299
	4.6%	4.0%	6.4%	4.1%	5.8%	5.5%	4.4%	4.4%	6.7%
D, cell division and chromosome partitioning	41	34	34	19	18	32	46	39	48
	0.9%	0.8%	0.7%	0.7%	2.8%	0.6%	0.6%	0.9%	1.1%
E, amino acid transport and metabolism	334	291	350	202	177	477	425	229	491
	7.5%	7.0%	8.1%	7.6%	9.5%	8.6%	5.4%	5.5%	11.1%
F, nucleotide transport and metabolism	79	82	87	69	49	101	102	63	85
	1.8%	1.9%	2.0%	2.6%	2.6%	1.8%	1.3%	1.5%	1.9%
G, carbohydrate transport and metabolism	205	289	367	95	160	223	539	217	242
	4.6%	7.0%	8.5%	3.6%	8.6%	4.0%	6.9%	5.2%	5.5%
H, coenzyme metabolism	152	106	123	66	47	150	172	115	164
	3.4%	2.5%	2.8%	2.5%	2.5%	2.7%	2.2%	2.7%	3.7%
I, lipid metabolism	118	88	83	72	24	195	213	109	162
	2.7%	2.1%	1.9%	2.7%	1.2%	3.5%	2.7%	2.6%	3.6%
J, translation, ribosomal structure, and biogenesis	168	243	258	211	178	326	205	162	171
	3.7%	5.9%	6.0%	8.0%	9.5%	5.9%	2.6%	3.9%	3.9%
K, transcription	270	289	280	118	73	447	713	187	392
	6.1%	7.0%	6.5%	4.4%	4.6%	8.0%	9.1%	4.5%	8.9%
L, DNA replication, recombination, and repair	143	133	220	119	87	140	233	252	240
	3.2%	3.2%	5.1%	4.5%	0.9%	2.5%	3.0%	6.0%	5.4%
M, cell envelope biogenesis, outer membrane	222	178	235	78	70	257	258	217	244
	5.0%	4.3%	5.4%	2.9%	3.7%	4.6%	3.3%	5.2%	5.5%
N, cell motility and secretion	255	54	107	11	56	141	68	183	177
	5.8%	1.3%	2.5%	0.4%	3.0%	2.5%	0.9%	4.4%	4.0%
O, Posttranslational modification, protein turnover, chaperones	134	98	128	89	52	182	159	148	158
	3.0%	2.3%	2.9%	3.3%	2.8%	3.3%	2.0%	3.5%	3.6%
P, inorganic ion transport and metabolism	159	161	191	81	69	293	195	187	233
	3.6%	3.9%	4.4%	3.0%	3.7%	5.3%	2.5%	4.5%	5.3%
Q, secondary metabolites biosynthesis, transport, and catabolism	130	88	68	44	18	173	290	122	181
	2.9%	2.1%	1.5%	1.6%	0.9%	3.1%	3.7%	2.9%	4.1%
R, general function prediction only	358	348	338	241	191	491	609	332	458
	8.0%	8.6%	7.9%	9.1%	10%	8.8%	7.8%	7.9%	10.4%
S, function unknown	250	308	309	220	130	459	299	209	329
	5.6%	7.4%	7.2%	8.3%	7.0%	8.2%	3.8%	5.0%	7.4%
T, Transduction mechanisms	304	121	134	75	50	233	390	194	345
	6.4%	2.9%	3.1%	2.8%	2.6%	4.2%	5.0%	4.6%	7.8%
Not in cogs	1162	1033	692	709	300	942	2564	1035	931
	24%	25%	16%	26%	16%	16.9%	32.8%	24.8%	17.4%
Total no. of ORFs	4431	4112	4279	2629	1858	5567	7825	4182	5350
Genome size, Mb	4.75	4.21	4.64	2.65	1.86	6.26	8.67	5.08	6.18
ORFs/100 kb	93.22	97.56	92.25	99.30	99.90	88.88	90.33	82.44	86.54

*cv, *C. violaceum*; bs, *Bacillus subtilis*; ec, *Escherichia coli*; dr, *Deinococcus radiodurans*; tm, *Thermotoga maritima*; pa, *P. aeruginosa*; sc, *Streptomyces coelicolor*; xcc, *Xanthomonas campestris citrus*; pp, *Pseudomonas putida*.

extent due to the presence or absence of particular ORFs within the genome, which is a reflection of the overall differential distribution of ORFs between free-living and commensal organisms. In contrast, the ORFs with highest similarity to *N. meningitidis* mostly belong to cog category J (ribosomal structure, biogenesis, and translation) and are present in all four genomes. This is in keeping with the concept that phylogenetic relationships are best reflected in ORFs for core housekeeping and structural proteins.

We undertook a survey of the general distribution of ORF functions using cog because it allows a standardized comparison with other sequenced genomes (see Table 2 and supplementary information at www.brgene.lncc.br/cviolaceum). This revealed that, in common with several of the other free-living bacteria, *C. violaceum* has a high proportion of ORFs associated with signal transduction mechanisms (cog category T) as well as cell motility and secretion (cog category N). These functions are directly involved in environmental interactions, and the larger number of ORFs in these categories thus reflects the need to be able to withstand environmental variability, which is not typically encountered by commensal organisms. We focused much of our attention during the analysis of the genome on understanding how the overall informational capacity of the genome, as illus-

trated by these tendencies, correlates with the ability of the organism to adapt to different environmental challenges.

General Metabolism. As expected for free-living organisms, the central and intermediary metabolic pathways present in *C. violaceum* include the synthesis and catabolism of all 20 amino acids as well as the purine and pyrimidine nucleotides. In addition, there are pathways for the synthesis of a wide range of cofactors and vitamins, although those leading to pantothenate and biotin are incomplete. Biosynthesis of complex polysaccharides including cellulose (but not glycogen) occurs as well as the synthesis and degradation of a variety of lipids used for energy supply, membrane formation, or energy storage including triacylglycerol, phospholipids, and lipopolysaccharide.

The ability of *C. violaceum* to thrive under diverse environmental conditions is clearly facilitated by its versatile energy-generating metabolism that is capable of exploiting a wide range of energy sources by using appropriate oxidases and reductases. These collectively permit both aerobic and anaerobic respiration (see supplementary information at www.brgene.lncc.br/cviolaceum). In the total absence of oxygen, nitrate or fumarate are used as final electron acceptors. The absence of nutrients also seems well tolerated through ORFs that act in response to

Brazilian National Genome Project Consortium

starvation conditions, many of which protect against oxidative damage. Examples include ORFs that respond to carbon starvation (*cstA*: CV0762 and CV1662) and those involved in peptide utilization (CV1098, CV1099, and CV1101) (30), the stringent starvation ORFs *sspA* and *sspB* (CV4005 and CV4004), which are induced by glucose, nitrogen, phosphate, or amino acid starvation (31), the DNA protection during prolonged starvation protein (Dps: CV4253), and the *pho* regulon.

Transporters. Transport-related membrane proteins mediate the bacterium's direct metabolic interactions with the complex soil and aquatic environments that it inhabits. We classified the 496 ORFs of this kind (≈11% of total ORF number) according to the Transport Protein Database, which reveals an extended collection of specific transporters (see supplementary information www.brgene.lncc.br/cviolaceum). The largest number of ORFs (212) are primary active transporters (class 3), of which 119 belong to the ATP-binding cassette transporter superfamily and 26 to the type III (virulence-related) pathway family. In addition, oxidoreduction-driven transporters are represented by 35 ORFs. Class 2, electrochemical potential-driven transporters, account for 154 ORFs, of which 144 are various kinds of porters, such as those of the major facilitator superfamily (MFS, 46 ORFs), the drug-metabolite transporter family (DMT, 13 ORFs), the resistance nodulation cell-division family (RND, 10 ORFs), the resistance-to-homoserine/threonine family (RhtB, 7 ORFs), and the C4-dicarboxylate uptake family (DCU, 2 ORFs). The presence of multidrug-resistance ORFs, belonging to four of the five families of drug exclusion translocases (32), illustrates the contribution of membrane transport systems to the capacity of *C. violaceum* to withstand environmentally unfavorable conditions. The transporters of heavy metals include *zntA* (CV1154), which provides *C. violaceum* with the potential for the bioremediation of xenobiotics. Also within class 2 are the ion gradient-driven energizers that are exclusively members of the TonB family (10 ORFs). There is a total of 35 ORFs related to iron metabolism, a particular priority for the bacterium, that include enterobactin, bacterioferritin, iron-storage proteins, and proteins for iron transport under anaerobic conditions in addition to the TonB-related proteins (33). The third most numerous class is the channels/pores (class 1), with 62 ORFs including 17 α-type channels and 41 β-barrel porins. Among the latter, there is one sugar porin and several outer membrane-linked receptors and factors. This class includes a number of transport systems that facilitate resistance to physical change. In this context, in addition to the ion transporters, there are systems that control the movement of other solutes across the bacterial cell membrane, as well as *aqpZ* (CV2864), which is selectively permeable to water (34). The four remaining classes, namely group translocators (class 4, 6 ORFs), transport electron carriers (class 5, 7 ORFs), accessory factors involved in transport (class 8, 25 ORFs), and incompletely characterized transport systems (class 9, 30 ORFs), comprise a total of 68 ORFs.

Stress Adaptation. The notable abundance of *C. violaceum* in the Rio Negro is indicative of its ability to simultaneously withstand a variety of relatively harsh environmental conditions including the scarcity of nutrients, high temperatures (often ≈40°C), high levels of radiation, and elevated concentrations of toxic agents including reactive oxygen species (2, 3 and 5). To a significant extent, the ability to cope with such environmental stress stems from the plethora of specific transporters present. Most crucially, these transporters permit the efficient exploitation of even very low concentrations of nutrients and are also responsible for the ability to withstand many toxic agents, although in the latter case several other types of resistance proteins are also operative. These include the organic hydroperoxide-resistance protein *ohr* (CV0209 and CV2493), disulfide oxidase *dsbA* (CV3998), and the alkylating

agents-inducible *aidB* (CV4136) as well as generic glutathione peroxidases, catalases, and aldolases (35). Specific protection against oxidative stress in *C. violaceum* is provided by the two major transcriptional regulators SoxR (CV2793) and OxyR (CV3378), and similar, hydrogen peroxide-inducible ORFs such as *dps* and *fur* and other ORFs are also present. A further crucial contribution to the resistance of environmental toxicity is provided by a series of proteins that ensure maintenance of cellular integrity. These include the OmlA lipoprotein (CV1796), also present in *P. aeruginosa* and *Burkholderia cepacia*, which provides resistance to anionic detergents and various antibiotics through the maintenance of cell envelope integrity under stress conditions (36, 37) as well as the mechanosensitive channel encoded by *mscL* (CV1360) that serves as an osmotic gauge (38).

Elevated temperatures are combated via a number of responses as indicated by the presence of 14 heat-shock-related ORFs including the DnaJ-DnaK-GrpE (Hsp70: CV1642, CV1643, and CV1645), the GroEL/GroES (*mopAB*) (CV3232, CV3233, CV4014, and CV4015), and the ClpA/B (CV1944, CV2557, CV2558, and CV3669) systems in addition to HscA/B cochaperones (CV1089 and CV1091), Hsp90 (HptG: CV1318), Hsp20 (CV1177), Hsp33 (CV2000), and Htpx (CV3109 and CV4263).Tolerance to UV radiation is provided by *uvrABC* (excinuclease/CV1893, CV3152, and CV1305) and *uvrD* (CV0205). In addition, however, there is evidence that violacein (CV3271 to CV3274) also contributes to protection against UV radiation (3).

The exquisite control of transcription that would be expected to be necessary bring the appropriate permutations of genes into play at any one time is effected by the combination of basic transcriptional mechanisms, such as RNA polymerase and common sigma factors, σ^{70} (*rpoD*), σ^{54} (*rpoN*), σ^{32} (*rpoH*), σ^{38} (*rpoS*), σ^{28} (*fliA*), σ^{24} (*rpoE*), and anti-σ^{28} factor (*flgM*), together with a large number of transcriptional activators and repressors that interact with alternative sigma factors involved in bacterial stress responses such as the 36 LysR, 14 AraC, 14 TetR, 12 Mar, 9 GntR, 5 Mer, 5 AsnC, 4 AsrR, 4 Crp/Fnr, 2 DeoR, 2 cold-shock, and 1 LacI family member ORFs.

Motility. An important contribution to the ability of *C. violaceum* to cope with environmental variability comes from its chemotactic capacity. A total of 68 ORFs are related to chemotaxis, of which 41 code for the methyl-accepting chemotaxis proteins. In comparison *P. aeruginosa* has a total of 43 chemotaxis-related ORFs (29), of which 26 are methyl-accepting chemotaxis proteins. Most chemotaxis-related ORFs are scattered throughout the genome, and none exhibit closest similarity with ORFs of the phylogenetically closely related *Neisseria* but rather with other free-living bacteria belonging mainly to the genera *Pseudomonas* (18 ORFs) and *Ralstonia* (10 ORFs). Some 64 ORFs related to flagellar structure and function were identified. The majority of these are contained in five operons (two *fli*, two *flg*, and one *flh*), although there are also several outlying ORFs for flagellar components (see supplementary information www.brgene.lncc.br/cviolaceum).

Quorum Sensing. Proteins that synthesize the specific autoinducers of quorum-sensing-controlled systems are evolutionarily well conserved and comprise the LuxR-LuxI family of transcriptional regulators (39). In *C. violaceum* two adjacent genes, *cviI* (CV4091) and *cviR* (CV4090), homologous to *luxI* and *luxR*, respectively, are transcribed from opposite strands and are convergently expressed with an overlap of 73 bp.

A number of *C. violaceum* phenotypic characteristics under quorum-sensing regulation have been reported including production of the purple pigment violacein (40), cyanide production (via the *hcnABC* operon), and degradation (11) through both the *cynT* (cyanate permease: CV1881) operon as well as *cynS* (cyanase: CV1880). ORFs coding for extracellular chitinases have also been

reported to be under quorum-sensing control (41). These ORFs are probably responsible for the ability of *C. violaceum* to survive on chitin as sole carbon and nitrogen source (42). Other ORFs present in *C. violaceum* reportedly controlled by quorum sensing (29) are those coding for elastase (*lasA* and *lasB*) and the antibiotic phenazine (CV0931 and CV2663). Furthermore, some genes coding for extracellular enzymes (for example, serine protease, collagenase, and oligopeptidase) exhibit upstream regulatory sequences homologous to those found in quorum-sensing-controlled genes and thus are possibly also regulated in this way.

Pathogenicity. Although *C. violaceum* is considered a saprophyte, it is also an occasional pathogen of human and animals with most cases of human infection occurring either early in childhood or in immunocompromised individuals (43). However, the fact that the Rio Negro is the source of drinking water for the population living around it, without there being widespread infection, indicates the low infectivity of this organism.

The lack of frequent human infection would be expected to select against the retention of purely pathogenesis-related genes. Thus, an unexpected finding was the presence of ORFs encoding type III secretory system (TTSS) components similar to those in *Salmonella typhimurium* (44) and *Yersinia pestis* (45). The TTSS is thought to be strictly associated with the infection of both animal or plant cells and acts as a molecular syringe for the secretion of effector molecules that provoke cytoskeletal rearrangements in the host cell (46). Because effectors with similarity to phytopathogen-associated genes (47) were not found, it seems unlikely that TTSS in *C. violaceum* plays a role in plant infection. Indeed, the similarity of the systems found to those in human pathogens suggests that they contribute to human infection. However, a detailed analysis of the *S. typhimurium*-like TTSS showed that some key ORFs including *invI* and *invH* [which have been demonstrated to play important roles in invasion (48, 49)] and *sicP* [a *Salmonella* invasion chaperone involved with the secretion of the tyrosine phosphatase SptP (50)] are absent in *C. violaceum*. The lack of these and other pathogenicity-related ORFs may account for the generally poor ability of the organism to infect humans. It is likely that the presence of these islands is isolate-specific. In PCR-based assays we found evidence for their presence in some isolates from natural Brazilian environments but not in others (see supplementary information at www.brgene.lncc.br/cviolaceum). The similarity of the two TTSSs with those found in other bacterial species, their presence in pathogenicity islands, and the fact that they are quite distinct from those found in the closely related opportunistic pathogen *P. aeruginosa* are all consistent with these ORFs being present in the *C. violaceum* genome due to recent lateral transfer.

Twelve ORFS encoding hemolysin-like proteins (CV0231, CV0360, CV0362, CV0513, CV0516, CV0656, CV1917, CV1918, CV2873, CV3275, CV3342, and CV4301) are found in both virulent and nonvirulent *C. violaceum* soil isolates (51). Type I and II secretory systems, both found in the *C. violaceum* genome, are likely to be also operative in free-living conditions despite their role as virulence factors in pathogenic bacteria (52, 53). The same holds true for genes coding for ubiquitous components of free-living Gram-negative bacteria (54, 55), which may also play a significant role in stimulating immune responses in the infected host such as the cell-wall-associated lipopolysaccharide and peptidoglycan.

Biotechnological Potential of *C. violaceum*. In addition to the operon responsible for the synthesis of the well studied violacein pigment (CV3274, CV3273, CV3272, and CV3271), there are many other ORFs encoding products of biotechnological and medical interest. For example, environmental detoxification may be mediated by an acid dehalogenase (CV0864), possibly active on xenobiotics or metabolic products (56), and also both by an operon for arsenic resistance (CV2438 and CV2440) and en-

zymes that catalyze the hydrolysis of cyanate (57). Conversely, cyanide can be used in gold recovery (18) besides being associated with the suppression of root fungi diseases (58). Of agricultural interest are the several chitinases (CV2935, CV3316, and CV4240) that are potential biocontrol agents against insects, fungi, and nematodes (59, 60). In addition, an insecticidal and nematocidal protein (CV1887) similar to those from *Xenorhabdus bovienii* and *Photorhabdus luminescens* (61) is also synthesized by *C. violaceum* and warrants further studies.

ORFs for two paraquat-inducible proteins (CV2547 and CV2548), potentially useful in bioengineering crops resistant to this herbicide, were found closely positioned in the genome. In addition, ORFs for the synthesis of medically relevant compounds include a polyketide synthase (CV4293) and other proteins applicable to antibiotic synthesis, genes for the synthesis of phenazine (CV0931 and CV2663) with potential antitumor activity, and hemolysins (CV0231, CV0513, CV1918, CV3342, and CV4301) with potential as anticoagulants. It is established already that *C. violaceum* has the capacity for the synthesis of polyhydroxyalkanoate polymers (18, 19), which have physical properties similar to propylene, making them an important renewable source of biodegradable plastic. In addition, we have now identified ORFs related to cellulose biosynthesis (CV2675, CV2677, and CV2678) that also might represent a valuable commodity, because bacterial cellulose differs from that produced by plants in its three-dimensional structure, degree of polymerization, and physicochemical properties (62).

Conclusions

The sequence and annotation data that we have generated reveal that the adaptability and versatility that *C. violaceum* exhibits depend on a large and complex genome containing a large proportion of ORFs that are specifically related to the ability of the organism to interact and respond to the environment. We also demonstrate that this genomic complexity might have practical importance in that it translates into the bacterium being an important potential source of biotechnologically exploitable genes. The identification of such genetic resources in *C. violaceum*, a free-living tropical bacteria, justifies the contemplation of strategic high-throughput programs to survey further the genomes of such organisms. Their inclusion in the pipeline that leads to the production of industrially useful genes, enzymes, and secondary metabolites would benefit not only the biotechnological and pharmaceutical industries in the developing world, where most tropical biodiversity is located, but would also provide a further stimulus to the preservation of the precious ecosystems where these organisms are found.

1. Boisbaudran, L. (1882) *Comp. Rend. Acad. Sci.* **94,** 562–562.
2. Caldas, L. R. (1990) *Cienc. Hoje* **11,** 55–57.
3. Caldas, L. R., Leitão, A. A. C., Santos, S. M. & Tyrrell, R. M. (1978) in *Proceedings of the International Symposium on Current Topics in Radiology and Photobiology*, ed. Tyrrell, R. M. (Academia Brasileira de Ciências, Rio de Janeiro), pp. 121–126.
4. Souza, A. O., Aily, D. C. G., Sato, D. N. & Durán, N. (1999) *Rev. Inst. Adolfo Lutz* **58,** 59–62.
5. Durán, N., Antonio, R. V., Haun, M. & Pilli, R. A. (1994) *World J. Microbiol. Biotechnol.* **10,** 686–690.
6. Leon, L. L., Miranda, C. C., Souza, A. O. & Durán, N. (2001) *J. Antimicrob. Chemother.* **48,** 449–450.
7. Lichstein, H. C. & van de Sand, V. F. (1945) *J. Infect. Dis.* **76,** 47–51.
8. Lichstein, H. C. & van de Sand, V. F. (1946) *J. Bacteriol.* **52,** 145–146.
9. Durán, N., Erazo, S. & Campos, V. (1983) *An. Acad. Bras. Cien.* **55,** 231–234.
10. Durán, N. (1990) *Cienc. Hoje* **11,** 58–60.
11. Duran, N. & Menck, C. F. (2001) *Crit. Rev. Microbiol.* **27,** 201–222.
12. Ueda, H., Nakajima, H., Hori, Y., Goto, T. & Okuhara, M. (1994) *Biosci. Biotechnol. Biochem.* **58,** 1579–1583.
13. Melo, P. S., Maria, S. S., Vidal, B. C., Haun, M. & Duran, N. (2000) *In Vitro Cell Dev. Biol. Anim.* **36,** 539–543.
14. Forsyth, W. G. C., Hayward, A. C. & Roberts, J. B. (1958) *Nature* **182,** 800–801.
15. Steinbüchel, A., Debzi, E. M., Marchessault, R. H. & Timm, A. (1993) *Appl. Microbiol. Biotechnol.* **39,** 443–449.
16. Gourson, C., Benhaddou, R, Granet, R., Krausz, P., Verneuil, B., Branland, P., Chauvelon, G., Tribault, J. F. & Saulnier, L. (1999) *J. Appl. Pollu. Sci.* **74,** 3040–3045.
17. Smith, A. D. & Hunt, R. J. (1985) *J. Chem. Technol. Biotechnol.* **35,** 110–116.
18. Campbell, S. C., Olson, G. J., Clark, T. R. & McFeters, G. (2001) *J. Ind. Microbiol. Biotechnol.* **26,** 134–139.
19. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomp, J. F., Dougherty, B. A. & Merrick, J. M. (1995) *Science* **269,** 496–512.
20. Hanke, J., Sanchez, D. O., Henriksson, J., Aslund, L., Pettersson, U., Frasch, A. C. & Hoheisel, J. D. (1996) *Biotechniques* **21,** 686–688, 690–693.
21. Carraro, D. M., Camargo, A. A., Salim, A. C., Grivet, M., Vasconcelos, A. T., Simpson, A. J. G. (2003) *Biotechniques* **34,** 626–628, 630–632.
22. Kanehisa, M. & Goto, S. (2000) *Nucleic Acids Res.* **28,** 29–34.
23. Tatusov, R., Galperin, M., Natale, D. & Koonin, E. (2000) *Nucleic Acids Res.* **28,** 33–36.
24. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., *et al.* (2000) *Bioinformatics* **16,** 1145–1150.
25. Nakai, K. (2000) *Adv. Protein Chem.* **54,** 277–344.
26. Francino, M. P. & Ochman, H. (1997) *Trends Genet.* **13,** 240–245.
27. Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J. C., Cattolico, L., *et al.* (2000) *Nature* **415,** 497–502.
28. Parkhill, J., Achtman, M., James, K. D., Bentley, S. D., Churcher, C., Klee, S. R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., *et. al.* (2000) *Nature* **404,** 502–506.
29. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., *et al.* (2000) *Nature* **406,** 959–964.
30. Schultz, J. E. & Matin, A. (1991) *J. Mol. Biol.* **218,** 129–140.
31. Williams, M. D., Ouyang, T. X. & Flickinger, M. C. (1994) *Mol. Microbiol.* **11,** 1029–1043.
32. Nikaido, H. (1996) *J. Bacteriol.* **178,** 5853–5859.
33. Faraldo-Gomez, J. D. & Sansom, M. S. (2003) *Nat. Rev. Mol. Cell Biol.* **4,** 105–116.
34. Calamita, G. (2000) *Mol. Microbiol.* **37,** 254–262.
35. Vergauwen, B., Pauwels, F., Vaneechoutte, M. & Van Beeumen, J. J. (2003) *J. Bacteriol.* **185,** 1572–1581.
36. Ochsner, U. A., Vasil, A. I., Johnson, Z. & Vasil, M. L. (1999) *J. Bacteriol.* **181,** 1099–1109.
37. Lowe, C. A., Asghar, A. H., Shalom, G., Shaw, J. G. & Thomas, M. S. (2001) *Microbiology* **147,** 1303–1314.
38. Moe, P. C., Blount, P. & Kung, C. (1998) *Mol. Microbiol.* **28,** 583–591.
39. Gray, K. M. & Garey, J. R. (2001) *Microbiology* **147,** 2379–2387.
40. McClean, K. H., Winson, M. K., Fish, L., Taylor, A., Chhabra, S. R., Camara, M., Daykin, M., Lamb, J. H., Swift, S., Bycroft, B. W., *et al.* (1997) *Microbiology* **143,** 3703–3711.
41. Chernin, L. S., Winson, M. K., Thompson, J. M., Haran, S., Bycroft, B. W., Chet, I., Williams, P. & Stewart, G. S. (1998) *J. Bacteriol.* **180,** 4435–4441.
42. Streischsbier, F. (1983) *FEMS Microbiol. Lett.* **143,** 3703–3711.
43. Richard, C. (1993) *Bull. Soc. Pathol. Exot.* **86,** 169–173.
44. Kimbrough, T. G. & Miller, S. I. (2002) *Microbes Infect.* **4,** 75–82.
45. Tyler, B. M. (2002) *Annu. Rev. Phytopathol.* **40,** 137–167.
46. Galan, J. E. & Collmer, A. (1999) *Science* **284,** 1322–1328.
47. Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., Churcher, C., Mungall, K. L., Bentley, S. D., Holden, M. T., *et al.* (2001) *Nature* **413,** 523–527.
48. Collazo, C. M., Kierler, M. K. & Galán, J. E. (1995) *Mol. Microbiol.* **15,** 25–38.
49. Watson, P. R., Paulin, S. M., Bland, P., Jones, P. W. & Wallis, T. S. (1995) *Infect. Immun.* **63,** 2743–2754.
50. Stebbins, C. E. & Galan, J. E. (2001) *Nature* **414,** 77–81.
51. Miller, D. P., Blevins, W. T., Steele, D. B. & Stowers, M. D. (1988) *Can. J. Microbiol.* **34,** 249–255.
52. Darzins, A. & Russell, M. A. (1997) *Gene* **192,** 109–115.
53. Tonjum, T. & Koomey, M. (1997) *Gene* **192,** 155–163.
54. Ingalls, R. R., Monks, B. G., Savedra, R., Jr., Christ, W. J., Delude, R. L., Medvedev, A. E., Espevik, T. & Golenbock, D. T. (1998) *J. Immunol.* **161,** 5413–5420.
55. Rietschel, E. T., Schletter, J., Weidemann, B., El-Samalouti, V., Mattern, T., Zahringer, U., Seydel, U., Brade, H., Flad, H. D. & Kusumoto, S., *et al.* (1998) *Microb. Drug Resist.* **4,** 37–44.
56. Janssen, D. B., Pries, F. & van der Ploeg, J. R. (1994) *Annu. Rev. Microbiol.* **48,** 163–191.
57. Anderson, P. M., Sung, Y. C. & Fuchs, J. A. (1990) *FEMS Microbiol. Rev.* **7,** 247–252.
58. Laville, J., Blummer, C., von Schroetter, C., Gaia, V., Défago, G., Keel, C. & Haas, D. (1998) *J. Bacteriol.* **180,** 3187–3196.
59. Cronin, D., Moenne-Loccoz, Y., Dunne, C. & O'Gara, F. (1997) *Eur. J. Plant Pathol.* **103,** 443–440.
60. Patil, R. S., Ghormade, V. & Despande, M. V. (2000) *Enzyme Microb. Technol.* **26,** 473–483.
61. Chen, G., Zhang, Y., Li, J., Dunphy, G. B., Punja, Z. K. & Webster, J. M. (1996) *J. Invertebr. Pathol.* **68,** 101–108.
62. Romling, U. (2002) *Res. Microbiol.* **153,** 205–212.

MICROBIOLOGY

Brazilian National Genome Project Consortium